Symplectic maps, variational principles, and transport

J. D. Meiss

Program in Applied Mathematics, University of Colorado, Boulder, Colorado 80309

Symplectic maps are the discrete-time analog of Hamiltonian motion. They arise in many applications including accelerator, chemical, condensed-matter, plasma, and fluid physics. Twist maps correspond to Hamiltonians for which the velocity is a monotonic function of the canonical momentum. Twist maps have a Lagrangian variational formulation. One-parameter families of twist maps typically exhibit the full range of possible dynamics–from simple or integrable motion to complex or chaotic motion. One class of orbits, the minimizing orbits, can be found throughout this transition; the properties of the minimizing orbits are discussed in detail. Among these orbits are the periodic and quasiperiodic orbits, which form a scaffold in the phase space and constrain the motion of the remaining orbits. The theory of transport deals with the motion of ensembles of trajectories. The variational principle provides an efficient technique for computing the flux escaping from regions bounded by partial barriers formed from minimizing orbits. Unsolved problems in the theory of transport include the explanation for algebraic tails in correlation functions, and its extension to maps of more than two dimensions.

CONTENTS

I. SYMPLECTIC MAPPINGS

A. Introduction

A dynamical system consists of a phase space describing the allowed states of a system and a rule defining the temporal evolution of those states. The evolution can be continuous, as for differential equations, or discrete, as for a mapping. Virtually every model of physical phenomena is a dynamical system; furthermore, most of the fundamental models of physics are Hamiltonian dynamical systems. The latter give rise to symplectic mappings. For example, the mapping defined by a Hamiltonian flow taking an initial condition to a state some finite time later is a symplectic map. Symplectic mappings are prominent in studies of charged-particle motion in particle accelerators, chemical reactions, or magnetic plasma confinement. Less appreciated is the fact that the motion of a fluid particle in an incompressible fluid is also Hamiltonian, even when the fluid motion itself is viscous. Mappings are useful because for many purposes they are easier to study than differential equations—certainly any numerical solution of a differential equation involves iteration of a map (and if the system is Hamiltonian, care should be taken to ensure that the map is symplectic). Mappings are also more general than differential equations.

Typical questions of physical interest include the long-time stability of orbits and the determination of the regions accessible to the motion. For example, in a particle accelerator, one would like to confine trajectories within the tunnel for something like 10^{10} revolutions. Direct simulation of a dynamical system for such periods is often impossible and, even if possible, is suspect due to the numerical errors induced—thus the need for basic theoretical results on stability. Another class of problems concerns transport, that is, the determination of the time for a group of trajectories to move from one region of phase space to another. Even if the system were not strictly stable, it could be stable in practice if the transport times were longer than the lifetime of the system—such is probably the case for the planetary motions in the solar system, though clearly not so for asteroids. Transport calculations enter as well into the theory of chemical reactions. For example, in the scattering problem $AB+C\rightarrow A+BC$, transport connects the regions of phase space corresponding to reactants and products, and quantities of interest are the reaction probabilities and rates. These could be computed statistically based on the volume of accessible phase space, but such calculations are often quantitatively incorrect due to dynamical obstructions to the motion—objects we call partial barriers. We discuss them in Secs. VIII and IX.

In this article our primary concern is the theory of symplectic twist mappings. The twist property is common in physical applications. It is fortuitous that the twist property also permits the use of powerful tools, both geometrical and analytical, resulting in a set of striking and fruitful theorems. The essence of the twist condition is that the canonical momentum variable represents a velocity on phase space—larger momentum implies that the configuration variable increases more rapidly. For example, for the free particle, the velocity is directly proportional to the momentum. We adopt the notation (x,y) for the phase-space coordinates, where y is this privileged momentum coordinate, and x is its conjugate configuration.

We review the theoretical results in Secs. III–VII. The first of these, the Kolmogorov-Arnol'd-Moser (KAM) theorem, is a perturbative result—it implies that most of the invariant tori of integrable twist mappings are preserved under perturbation. The rest of the results we discuss are nonperturbative—they hold for any twist mapping. The proof of Birkhoff's theorem, Sec. IV, typifies the geometrical reasoning allowed by the twist condition. One consequence of this theorem is a nonexistence criterion for invariant circles of 2D twist maps. We next discuss analytical results that are based on the variational principle for twist maps.

The variational principle for twist maps is analogous to the Lagrangian-action formulation of analytical mechanics. Orbits are stationary points of the action function. What is most interesting about twist maps is that special extrema of the action, the minima and minimax points, lead to a class of orbits that are of great importance. These orbits each have a definite rotation frequency and satisfy ordering properties. For rational frequencies these orbits are the elliptic and hyperbolic orbits that form the island chains (see Sec. II). For irrational frequencies they are either invariant circles, when the system is weakly perturbed, or invariant Cantor sets—the cantori—when the system is strongly perturbed (see Sec. VII). Thus we obtain a general picture of the regular part of the phase space of these maps.

Chaotic orbits must wend their way through the obstacle course formed by the minimizing and minimax orbits

and their stable and unstable manifolds. We use these manifolds to construct partial barriers, and in Secs. VIII and IX, develop a theory of transport based on the flux of trajectories through these barriers. This theory is in direct contrast with a more uniform statistical picture of chaos and shows that chaotic zones must be partitioned into subsets that are often separated by effective barriers. One of the predictions of this theory is that any system with regular regions should exhibit slow, nonexponential decay of correlation functions. Another is a universal exponent for the onset of transport when an invariant circle is destroyed by perturbation, becoming a cantorus.

In the remainder of this section we review Hamiltonian dynamics, discuss the nature of symplectic flow, and provide examples.[1] We then define twist mappings, which will be our major concern. Those readers with little interest in the physical motivations for studying symplectic mappings can proceed directly to Sec. I.E, where twist maps are introduced.

B. Hamiltonian flows

A Hamiltonian flow is described by a function $H(\mathbf{p},\mathbf{q},t)$ and a set of differential equations

$$\frac{dq^i}{dt}=\frac{\partial H}{\partial p^i},\quad \frac{dp^i}{dt}=-\frac{\partial H}{\partial q^i}\ . \tag{1.1}$$

Here the q^i represent *configuration* coordinates and the p^i represent *canonical momenta*, $i=1,2,\ldots N$, for a system with N *degrees of freedom*. For example, $H=\frac{1}{2}\mathbf{p}^2+V(\mathbf{q})$ represents the energy of a set of particles interacting through a potential V. More compactly (and more generally), these equations can be written as

$$\dot{z}=\{z,H\}\ . \tag{1.2}$$

Here we use the symbol z to denote arbitrary coordinates on phase space regardless of its dimension—it is hoped that no confusion between scalars and vectors will arise. The coordinates are denoted by $z^m, m=1,2,\ldots 2N$, and $\{,\}$ represents the Poisson bracket. The latter is defined for any two functions $f(z)$ and $g(z)$ as

$$\{f,g\}\equiv\sum_{m,n=1}^{2N}\frac{\partial f}{\partial z^m}J^{mn}\frac{\partial g}{\partial z^n}=\sum_{k=1}^{N}\frac{\partial f}{\partial q^k}\frac{\partial g}{\partial p^k}-\frac{\partial g}{\partial q^k}\frac{\partial f}{\partial p^k}\ , \tag{1.3}$$

where J, the Poisson matrix, is the $2N\times 2N$ antisymmetric matrix

$$J=\begin{bmatrix}0 & I\\ -I & 0\end{bmatrix} \tag{1.4}$$

in (q^i,p^i) coordinates. In fact, J transforms as a contravariant tensor, and Hamilton's equations (1.2) are covariant.

Hamiltonian flow can be obtained from a variational principle. Consider a trial trajectory or path $\{\mathbf{q}(t),\mathbf{p}(t);t_0<t<t_1\}$ in phase space connecting the point $(\mathbf{q}_0,\mathbf{p}_0)$ to $(\mathbf{q}_1,\mathbf{p}_1)$. The *action* is a functional on such a path, defined as

$$S=\int_{t_0}^{t_1}[\mathbf{p}\cdot\dot{\mathbf{q}}-H(\mathbf{p},\mathbf{q},t)]\,dt\ . \tag{1.5}$$

Hamilton's principle states that the true path between the fixed end points $\mathbf{q}(t_0)=\mathbf{q}_0$ and $\mathbf{q}(t_1)=\mathbf{q}_1$ is one for which S is stationary:

$$0=\delta S=\int_{t_0}^{t_1}dt\left[\delta\mathbf{p}\cdot\dot{\mathbf{q}}+\mathbf{p}\cdot\delta\dot{\mathbf{q}}-\frac{\partial H}{\partial\mathbf{p}}\cdot\delta\mathbf{p}-\frac{\partial H}{\partial\mathbf{q}}\cdot\delta\mathbf{q}\right]\ . \tag{1.6}$$

Since the path is arbitrary in the phase space, the variations $\delta\mathbf{p}$ and $\delta\mathbf{q}$ are independent. Thus the coefficient of each must be zero. The coefficient of $\delta\mathbf{p}$ yields directly the equation of motion for $\mathbf{q}$. Integration by parts on $\delta\mathbf{q}$, and using the fixed end-point conditions on $\mathbf{q}$, gives the equation of motion for $\mathbf{p}$.

The action (1.5) is covariant and thus can be written in arbitrary coordinate systems.

The action principle is handy because it represents the equations of motion in a compact, scalar form; however, it also has more applications. In fact, a major theme of this paper is that the action can be used to compute quantities of physical importance. We shall first use the action principle to show that Hamiltonian flow is symplectic.

C. Symplectic mappings

A *mapping* is a transformation of each point in the phase space

$$z'=T(z)\ . \tag{1.7}$$

We shall consider only diffeomorphisms, that is, one-to-one mappings that are smooth and have smooth inverses. A function is of class C^n if it has n continuous derivatives. A C^0 diffeomorphism is also called a homeomorphism. Some of the results discussed here are valid for homeomorphisms, but most require some degree of differentiability.

An *orbit* is a sequence

$$\{\ldots,z_t,z_{t+1},\ldots\} \tag{1.8}$$

such that $z_{t+1}=T(z_t)$.

As we shall see below, mappings arise naturally from flows. An example is the transformation of phase space given by integrating every point forward one unit in time;

[1]For a more complete discussion of some of the topics in this section, consult Lichtenberg and Lieberman (1982) or, for the mathematically inclined, Arnol'd (1978), MacKay and Meiss (1987), or Arrowsmith and Place (1990).

another is the "return map." We shall discuss next the mappings that come from Hamiltonian flows. Such maps are termed "symplectic."

1. Integral invariant

We show here that the action of a loop is an invariant for Hamiltonian flow—the Poincaré integral invariant. The loops we consider are closed curves in the extended phase space $(\mathbf{q},\mathbf{p},t)$. One could, for example, choose a loop at some fixed time—a loop in ordinary phase space; however, the loop could just as well depend on time. We denote by the symbol $\mathcal{L}$ a loop that is contractible to a point. In terms of some parameter λ, the loop is given by $\{\mathbf{q}(\lambda),\mathbf{p}(\lambda),t(\lambda);0\le\lambda\le 1\}$. The action of $\mathcal{L}$ is the loop integral

$$S[\mathcal{L}]=\int_0^1\left[\mathbf{p}\cdot\frac{d\mathbf{q}}{d\lambda}-H\frac{dt}{d\lambda}\right]d\lambda$$
$$=\oint_{\mathcal{L}}\mathbf{p}\cdot d\mathbf{q}-H\,dt\ . \tag{1.9}$$

Here the second integral is a convenient notation for the first. A more compact notation for Eq. (1.9) is obtained by defining the vector $\mathbf{A}=(\mathbf{p},\mathbf{0},-H)$, and a line element $d\mathbf{l}=(d\mathbf{q},d\mathbf{p},dt)$, to give

$$S[\mathcal{L}]=\oint_{\mathcal{L}}\mathbf{A}\cdot d\mathbf{l}\ . \tag{1.10}$$

Every point on $\mathcal{L}$ constitutes an initial condition for Hamilton's equations, and we can evolve the loop by integrating from each point. This gives a two-dimensional tube $\mathcal{T}$, Fig. 1. Now consider any loop $\mathcal{L}'$ on $\mathcal{T}$ that is homotopically equivalent to $\mathcal{L}$ (i.e., $\mathcal{L}'$ must be obtained by sliding $\mathcal{L}$ along $\mathcal{T}$ in some continuous but otherwise arbitrary way; however, $\mathcal{L}'$ need not be a loop that is obtained by evolving $\mathcal{L}$ forward for a fixed time step). We wish to show that the action of $\mathcal{L}'$ is equal to $S[\mathcal{L}]$. The difference between $S[\mathcal{L}]$ and $S[\mathcal{L}']$ is the difference between the two loop integrals of the vector $\mathbf{A}$. These two loops bound a piece of the tube $\mathcal{T}$, and because $\mathcal{L}'$ is homotopic to $\mathcal{L}$ this piece of $\mathcal{T}$ is simply connected. Stokes's theorem[2] implies that the difference between the actions is equal to the integral of $\nabla\times\mathbf{A}$ over this piece of $\mathcal{T}$:

$$S[\mathcal{L}]-S[\mathcal{L}']=\int_{\mathcal{T}}\nabla\times\mathbf{A}\cdot d^2s\ , \tag{1.11}$$

where d^2s is the surface area element. A simple calculation shows that $\nabla\times\mathbf{A}=(-\partial H/\partial\mathbf{p},\partial H/\partial\mathbf{q},-1)$, which is in fact the negative of the velocity vector in extended phase space: $\nabla\times\mathbf{A}=-(\dot{\mathbf{q}},\dot{\mathbf{p}},1)$. By construction the velocity vector lies along $\mathcal{T}$, perpendicular to d^2s; so the integrand in Eq. (1.11) is zero, and

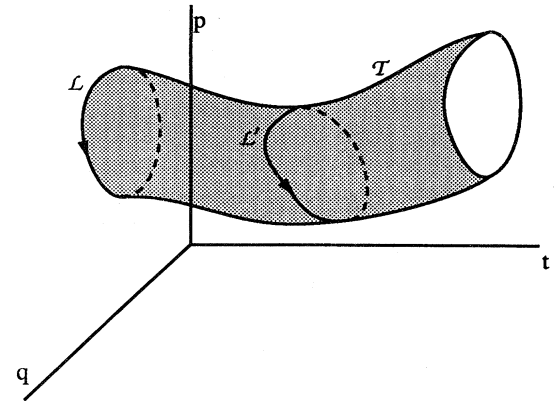

FIG. 1. Preservation of the loop action for Hamiltonian flows.

$$S[\mathcal{L}']=S[\mathcal{L}]\ . \tag{1.12}$$

A map that preserves the loop action is *symplectic.*

As an example, suppose H is independent of time; then it is constant along the motion given by (1.1). Consider any loop $\mathcal{L}$ contained within an energy surface $H=E$. Since H is constant on $\mathcal{L}$, it can be removed from the loop integral, and $\oint dt=0$, therefore

$$S[\mathcal{L}]=\oint_{\mathcal{L}}\mathbf{p}\cdot d\mathbf{q}\ \text{ (on an energy surface)}\ . \tag{1.13a}$$

The integral (1.13a) is the *symplectic area*; its value is the sum of the N areas of the projections of $\mathcal{L}$ on the canonical planes (q^i,p^i), shown in Fig. 2. Thus invariance of the action implies that the symplectic area is conserved along the flow of a time-independent Hamiltonian. It is important for applications that the loop $\mathcal{L}'$ need not be a loop obtained from $\mathcal{L}$ by evolving for a fixed time step (see Sec. I.C).

As a second application consider loops on fixed time

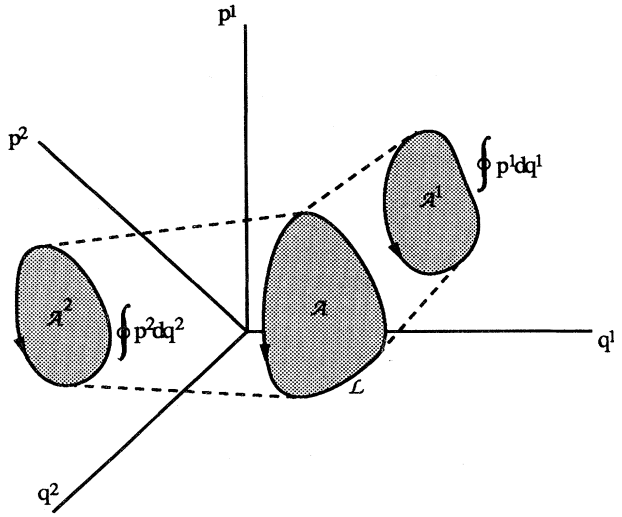

FIG. 2. Definition of the Poincaré integral invariant. The value of each of the projected symplectic areas shown is negative because the loops are traversed counterclockwise.

[2]Actually we are using the generalization of Stokes's theorem to many dimensions; the curl is generalized to the exterior derivative dA (Appendix A). The net result is the same as Eq. (1.11).

slices, $t(\lambda)=$constant. Then, even though H may depend upon time, $\oint_{\mathcal{L}} H\,dt=0$, and the action is

$$S[\mathcal{L}]=\oint_{\mathcal{L}}\mathbf{p}\cdot d\mathbf{q}\quad(\text{for } t=\text{constant})\ . \tag{1.13b}$$

Furthermore, Eq. (1.12) implies that the value of this action is the same for any loop $\mathcal{L}'$ that is on a constant t surface. Thus integrating Hamilton's equations forward by a fixed time step conserves the action (1.13b).

Most of our applications will deal with the special cases represented by Eq. (1.13).

2. Symplectic form

Using Stokes's theorem in reverse provides an alternative representation for (1.13)—it becomes the sum of integrals over the two-dimensional disks bounded by the projection of $\mathcal{L}$ onto each canonical plane (see Fig. 2):

$$S[\mathcal{L}]=\int_{\mathcal{A}} d\mathbf{p}\wedge d\mathbf{q}\equiv\sum_{i=1}^{N}\sigma_i\int_{\mathcal{A}^i}dp^i dq^i\ . \tag{1.14}$$

Here the wedge product represents an oriented area, so that σ_i is $+1$ if the projection of $\mathcal{L}$ is traversed clockwise or -1 if traversed counterclockwise in the canonical plane.[3]

For example, consider a loop that is a parallelogram with sides made from two vectors $\delta\bar{z}$ and δz, sketched in Fig. 3. Its symplectic area is the sum of each of the areas of its projections. We denote this ω

$$\omega(\delta z,\delta\bar{z})=\delta\mathbf{p}\cdot\delta\bar{\mathbf{q}}-\delta\mathbf{q}\cdot\delta\bar{\mathbf{p}}=\delta z^i\omega_{ij}\delta\bar{z}^j\ . \tag{1.15}$$

The antisymmetric form ω is called the *symplectic form.* In (q,p) coordinates, it is represented by the matrix

$$\omega=\begin{bmatrix}0 & -I\\ I & 0\end{bmatrix}, \tag{1.16}$$

which is the inverse of the Poisson tensor J.

3. Locally symplectic mappings

We can use the symplectic form to obtain a differential statement of the symplectic condition. Suppose T is a symplectic mapping; by definition T preserves the loop action (1.12). Consider an infinitesimal parallelogram at the point z, which is made from two arbitrary vectors δz and $\delta\bar{z}$. Under the mapping this parallelogram has an image at z'; each of the sides are given by the derivative of the mapping (1.7) at z:

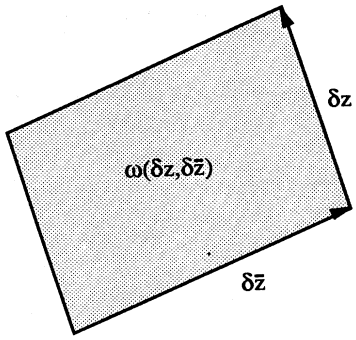

FIG. 3. Interpretation of the symplectic two-form.

$$\delta z'=DT(z)\,\delta z\equiv\left[\frac{\partial}{\partial z}T(z)\right]\delta z\ . \tag{1.17}$$

Thus the image is also a parallelogram. Equation (1.12) implies that the symplectic area of the image is equal to its initial value:

$$\omega(\delta z',\delta\bar{z}')=\omega(\delta z,\delta\bar{z})\ . \tag{1.18}$$

Using the definition (1.15), it is easy to see that this implies

$$\tilde{M}\omega M=\omega\ , \tag{1.19}$$

where ω denotes the matrix (1.16) and M denotes the Jacobian matrix

$$M^i_j\equiv(DT)^i_j=\frac{\partial z'^i}{\partial z^j}\ . \tag{1.20}$$

Equation (1.20) is the local requirement on the mapping T imposed by the integral invariant. Any map whose derivative satisfies (1.19) everywhere is *locally symplectic.* If the phase space is not simply connected, then the conservation of the integral invariant (1.13) for curves that cannot be deformed to a point is an additional requirement (see Sec. V.B). Maps that are symplectic in this second sense are *exactly* symplectic.

4. Reflexivity and volume preservation

A simple consequence of Eq. (1.19) follows from taking its determinant:

$$\mathrm{Det}(\tilde{M}\omega M)=\mathrm{Det}(\omega)\Longrightarrow(\mathrm{Det}M)^2=1\ ,$$

since $\mathrm{Det}(\omega)\neq 0$. This implies that $\mathrm{Det}(M)$ must be either ± 1. In fact, we shall show that $\mathrm{Det}(M)=+1$, and that any symplectic map is therefore volume and orientation preserving. In showing this we shall also obtain an important property of the eigenvalues of symplectic matrices.

Consider the eigenvalue problem for M. The characteristic equation is the ($2N$th)-order polynomial

$$\mathrm{Det}(M-\lambda I)=0\ . \tag{1.21}$$

[3]Unfortunately this is the reverse of a common convention; however, since we wish to have (q,p) represent horizontal and vertical coordinates, respectively, it seems that a minus sign must appear at some point.

Because the mapping is real, the characteristic polynomial is also real;

$$\lambda \text{ is an eigenvalue of } M \Longrightarrow \lambda^* \text{ is an eigenvalue of } M \ . \tag{1.22}$$

More interestingly, using Eq. (1.19) we can rewrite (1.21) as

$$\begin{aligned} 0 &= \mathrm{Det}(\omega)\,\mathrm{Det}(M-\lambda I) = \mathrm{Det}(\omega M - \lambda\omega) \\ &= \mathrm{Det}(\tilde{M}^{-1}\omega - \lambda\omega) = \mathrm{Det}(\tilde{M}^{-1} - \lambda I)\,\mathrm{Det}(\omega) \\ &= \mathrm{Det}(M^{-1} - \lambda I) \ . \end{aligned}$$

Thus if λ is an eigenvalue of M, it is also an eigenvalue of M^{-1}. Alternatively,

$$\lambda \text{ is an eigenvalue of } M \Longrightarrow \lambda^{-1} \text{ is an eigenvalue of } M \ . \tag{1.23}$$

Thus the characteristic polynomial is *reflexive*: it can be written in the form

$$\lambda^N + A\lambda^{N-1} + B\lambda^{N-2} + \cdots + B\lambda^{2-N} + A\lambda^{1-N} + \lambda^{-N} = 0 \ .$$

Since $\mathrm{Det}(M)$ is the product of its eigenvalues, (1.22) and (1.23) imply directly that

$$\mathrm{Det}(M) = 1 \ . \tag{1.24}$$

Thus two-dimensional symplectic maps, for example, preserve the oriented area element $dp_1 \wedge dq_1$. Conversely, any two-dimensional map that preserves area and orientation is locally symplectic.

Equations (1.22) and (1.23) imply that eigenvalues appear either in pairs or in quadruplets (Fig. 4). If λ is real, then it has a partner λ^{-1}. If λ is complex and has only one partner under (1.22) and (1.23), then $\lambda^* = \lambda^{-1}$; so it is on the unit circle. Furthermore, if $\lambda = 1$ is an eigenvalue, then it must have even multiplicity, since the phase space is even dimensional. Finally, if λ is neither real nor of unit modulus, then there must be a quadruplet of eigenvalues

$$\lambda, \lambda^{-1}, \lambda^*, \lambda^{-1*} \ . \tag{1.25}$$

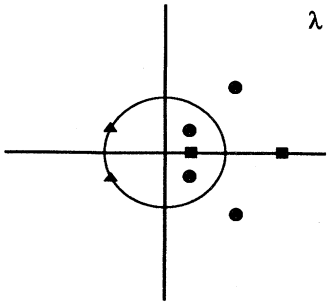

FIG. 4. Possible eigenvalues for a symplectic matrix in the complex plane. The triangles are a unit modulus pair, the squares are a real pair, and the circles are a quadruplet.

Of course, this case can occur only for four or more dimensions.

D. Return mappings

Consider a time-independent Hamiltonian. Since the energy is conserved, the flow occurs on a $(2N-1)$-dimensional energy surface $\mathcal{E}$ corresponding to a value $E = H$. Now suppose there is another $(2N-1)$-dimensional surface $\mathcal{Q}$ that is transverse (i.e., nowhere parallel) to the flow in some local region (see Fig. 5). The *Poincaré section* $\mathcal{S}$ is the $(2N-2)$-dimensional intersection of $\mathcal{E}$ with $\mathcal{Q}$. The *return mapping*, denoted $z' = T(z)$, is the function that takes an initial condition z on $\mathcal{S}$ to the point z' at which it first returns on $\mathcal{S}$. The Poincaré recurrence theorem states that if the energy surface is bounded (compact), almost all trajectories (all but a set of zero volume) that begin on $\mathcal{S}$ will eventually return to $\mathcal{S}$ (Cornfeld *et al.*, 1982). The return map is symplectic with action (1.13).

For example, let $\mathcal{Q}$ be the surface $q_N = \text{constant}$. It is transverse to the flow if

$$\frac{dq_N}{dt} = \frac{\partial H}{\partial p_N} \neq 0 \quad \text{on } \mathcal{Q} \ . \tag{1.26}$$

The Poincaré section $\mathcal{S}$ can be described by the coordinates $(q_1, p_1, \ldots, q_{N-1}, p_{N-1})$, since, with a choice of value for the energy, transversality (and the implicit function theorem) implies that $H(q_1, p_1, \ldots, q_N, p_N) = E$ can be inverted to obtain

$$p_N = p_N(q_1, p_1, \ldots, q_{N-1}, p_{N-1}; q_N, E) \ . \tag{1.27}$$

The return mapping T is parametrized by the choice of E and q_N. In this coordinate system, the action (1.13) reduces to $S = \sum_{i=1}^{N-1} \oint p^i dq^i$.

In the particular case of a two-degree-of-freedom Hamiltonian, $N = 2$, the mapping T acts on the two-

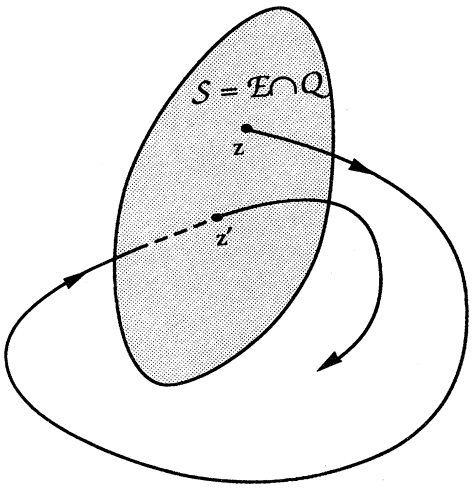

FIG. 5. Return mapping.

dimensional phase space (q_1,p_1). In this paper we shall almost always consider this two-dimensional case. There are many examples of physical interest, and we give two below.

1. Hénon-Heiles Hamiltonian

The Hénon-Heiles model (Hénon and Heiles, 1964) is a two-degree-of-freedom system with the Hamiltonian

$$H=\tfrac{1}{2}(p_x^2+p_y^2+x^2+y^2+2x^2y-\tfrac{2}{3}y^3) .$$

It was chosen to model the motion of a star in a galaxy with an axisymmetric distribution of matter. The Hamiltonian has a bounded energy surface when $E\le\frac{1}{6}$. The original pictures of the flow of this system were obtained using the surface $\mathcal{Q}$ defined by $x=0$, which is transverse to the flow for $p_x\neq 0$. Typically, one chooses $p_x>0$ to fix the branch of the function $p_x(y,p_y;E,x)$. Since $p_x^2\ge 0$, the domain of the mapping is restricted to the region $p_y^2+y^2-\frac{2}{3}y^3\le 2E$, which looks like an oval for E small and has a corner when $E=\frac{1}{6}$.

Though this Poincaré section is commonly used, any choice of a transverse surface $\mathcal{Q}$ will give a symplectic mapping; and since the Hamiltonian flow provides a smooth connection between various transverse surfaces, the structure of the mappings will be the same.

2. Passive tracers and magnetic fields

Volume-preserving flow in three dimensions also can be thought of as a Hamiltonian system and reduced to an area-preserving mapping, providing there are no null points of the flow. For example, consider an incompressible fluid with velocity field $\mathbf{v}(\mathbf{x})$, or a magnetic field $\mathbf{B}(\mathbf{x})$. The equations for the Lagrangian particle trajectories govern the motion of a passive tracer in the fluid. An understanding of the Hamiltonian nature of these equations is important for the study of mixing (Aref, 1984; Khakhar *et al.*, 1986; Ottino, 1989). Similarly, the equations for the magnetic-field lines are, to the lowest approximation, the equations of charged particles in small gyroradius orbits. This is especially applicable to magnetic confinement of plasmas. There are many applications of the study of such equations (Rosenbluth *et al.*, 1966; Dragt and Finn, 1976; Rechester and Rosenbluth, 1978; Chirikov, 1979a; Mynick and Krommes, 1980; Boozer and White, 1982). We shall use notation appropriate to the magnetic-field case.

The relevant equations take the form

$$\frac{d\mathbf{x}}{dt}=\mathbf{B}(\mathbf{x}) ,$$

where t is a parameter-measuring distance along the field lines (or streak lines). Whenever the magnetic field is nonvanishing, the system of equations (1.28) is Hamiltonian—in fact, it is equivalent to a one-degree-of-freedom, time-dependent Hamiltonian. Thus three-dimensional physical space is equivalent to the extended phase space (q,p,t) (Cary and Littlejohn, 1983). The action principle (1.5) can be shown to become

$$S=\int_{\mathbf{x}_0}^{\mathbf{x}_1}\mathbf{A}\cdot d\mathbf{l} , \tag{1.28}$$

where $\mathbf{A}(\mathbf{x})$ is the vector potential. In a general coordinate system, the equations of motion generated by Eq. (1.28) are noncanonical in form. There is an important special case that is naturally canonical—a toroidal system with coordinates (ψ,θ,ζ), where θ and ζ are the poloidal and toroidal angle variables, and for which the toroidal component of $\mathbf{B}$ does not vanish. One can show that a suitable radial coordinate ψ and a suitable gauge can be found so that

$$\mathbf{A}=\psi\nabla\theta-\chi\nabla\zeta . \tag{1.29}$$

The corresponding field is $\mathbf{B}=\nabla\psi\times\nabla\theta-\nabla\chi\times\nabla\zeta$. We have assumed that the contravariant component of $\mathbf{B}$ in the toroidal direction does not vanish:

$$B^{\zeta}=B\cdot\nabla\zeta=\nabla\psi\cdot\nabla\theta\times\nabla\zeta\neq 0 ;$$

this is equivalent to (ψ,θ,ζ) being a nonsingular coordinate system. Using Eq. (1.29), we see that the action (1.28) becomes

$$S=\int\psi\,d\theta-\chi(\psi,\theta,\zeta)\,d\zeta .$$

Comparing this with Eq. (1.5) shows that (θ,ψ) are a canonical pair of variables, and χ acts as the Hamiltonian with ζ playing the role of time. Periodicity in ζ implies that we can use the Poincaré section technique to construct an area-preserving mapping $T:(\psi,\theta)\to(\psi',\theta')$.

In general, since the flow is Hamiltonian at any point for which $\mathbf{B}\neq 0$, the two eigenvalues of the map satisfy $\lambda_1\lambda_2=1$, according to (1.22) and (1.23). The flow has a third eigenvalue that corresponds to the direction of $\mathbf{B}$; it must be 1 because the flow is volume preserving. This need not hold at null points of the flow—there the only restriction is $\lambda_1\lambda_2\lambda_3=1$.

E. Twist mappings

We now restrict consideration to two-dimensional maps and assume that the phase space (x,y) is a cylinder, with x being the angle coordinate. Such a phase space arises naturally in many examples, where y represents a momentum and so is unbounded, but x represents the angle coordinate of, for example, an oscillator. Let $T:(x,y)\to(x',y')$ be a symplectic map from the cylinder to itself, and suppose T is differentiable. Then T is a *twist* map (with twist to the right) if there is a K such that

$$\left.\frac{dx'}{dy}\right|_x\ge K>0 , \tag{1.30}$$

which means that x' is a monotonically increasing function of y. This is illustrated in Fig. 6—the first iterate of a vertical line ($x=$constant) tilts to the right (is a graph

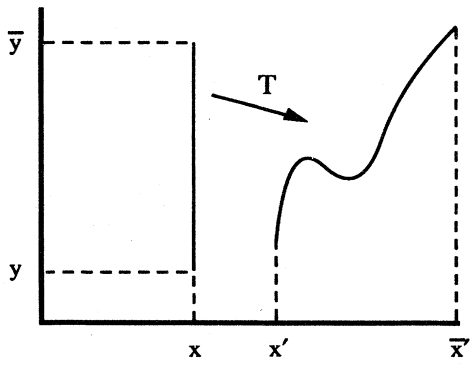

FIG. 6. Geometrical interpretation of the twist condition (1.30).

over x). The twist condition is natural physically, since y represents a momentum, and larger momentum usually implies larger velocity. Thus points with larger y should move farther in x. As noted in Fig. 6, this relation does not imply that $y'(x,y)$ is a function of y.

Since the map is differentiable, we can consider its action on a tangent vector $(\delta x,\delta y)$, as in (1.17):

$$\begin{bmatrix}\delta x'\\ \delta y'\end{bmatrix}=\begin{bmatrix}\frac{\partial x'}{\partial x} & \frac{\partial x'}{\partial y}\\ \frac{\partial y'}{\partial x} & \frac{\partial y'}{\partial y}\end{bmatrix}\begin{bmatrix}\delta x\\ \delta y\end{bmatrix}=M\begin{bmatrix}\delta x\\ \delta y\end{bmatrix}. \tag{1.31}$$

According to Eq. (1.24) the matrix M has unit determinant. The inverse of the linear map is represented by the derivative of T^{-1} as well as the inverse of M; thus

$$\begin{bmatrix}\frac{\partial x}{\partial x'} & \frac{\partial x}{\partial y'}\\ \frac{\partial x}{\partial x'} & \frac{\partial y}{\partial y'}\end{bmatrix}=M^{-1}=\begin{bmatrix}\frac{\partial y'}{\partial y} & -\frac{\partial x'}{\partial y}\\ -\frac{\partial y'}{\partial x} & \frac{\partial x'}{\partial x}\end{bmatrix}. \tag{1.32}$$

Therefore the twist condition implies that

$$\left.\frac{\partial x}{\partial y'}\right|_{x'}=-\left.\frac{\partial x'}{\partial y}\right|_{x}\le -K\ ; \tag{1.33}$$

so if T is a twist map, then T^{-1} is also a twist map, but one that twists to the left. Note that T^2 is not necessarily a twist map, and indeed typically is not, because the tilted line can rotate enough on the second iterate to violate the twist condition (T^2 is a member of a more general class of maps, called "tilt" maps, to which we shall refer in Sec. IV.C).

This paper will almost entirely concentrate on the study of area-preserving twist maps. The theory behind these maps is well developed, and the twist condition permits the proof of several important theorems. Moreover, twist maps occur commonly in applications.

F. Examples of twist maps

1. The cyclotron

Symplectic maps arise often in the study of particle accelerators (Carrigan *et al.*, 1982; Evans, 1983; Jowett *et al.*, 1986). The simplest accelerator is the cyclotron, which, though it is not a good example of modern design, provides a nice example of a twist map. Our model cyclotron consists of a constant magnetic field $\mathbf{B}=B_0\mathbf{e}_z$ and a time-dependent voltage drop $V\sin\omega t$ across a narrow azimuthal gap (Fig. 7).

Suppose there is an orbiting electron in the cyclotron. The time for an electron to go around one circuit of the cyclotron is

$$T=\frac{2\pi}{\Omega_c}=2\pi\frac{m\gamma c}{eB}=2\pi\frac{E}{eBc}\ , \tag{1.34}$$

where E is the particle energy $m\gamma c^2$, and γ is the relativistic factor. The change in energy upon traversing the gap is $\Delta E=-eV\sin\omega t$. Let (E,t) be the energy and time just before the electron reaches the gap; then after one circuit their new values are

$$E'=E-eV\sin\omega t,\quad t'=t+(2\pi/ceB)E'\ , \tag{1.35}$$

providing the kick is too small to reverse the velocity.

Defining normalized variables $y=\omega E/ceB=\omega/\Omega_c$, $x=\omega t/2\pi$, and $k=2\pi\omega V/cB$, Eq. (1.35) becomes the "standard map"

$$T:\begin{cases}y'=y-\frac{k}{2\pi}\sin(2\pi x)\ ,\\ x'=x+y'\ .\end{cases} \tag{1.36}$$

It depends on a single parameter, k, representing the strength of the nonlinear kick. It is important that $y'(x,y)$ appears in the second equation, so that the map preserves area. In the case discussed here, y' represents the energy after the kick and is therefore the proper value to use for calculating the next period. Since the map is taken at a fixed value of the angular position, the action (1.9) reduces to $-\oint H\,dt$ for this map; thus we should expect that (x,y) representing time and energy are appropriate canonical coordinates. The standard map has twist; in fact, $\partial x'/\partial y=1$.

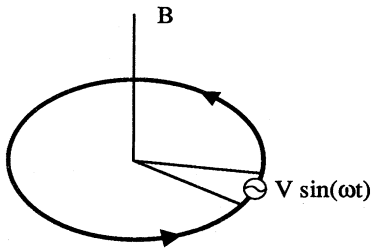

FIG. 7. Model cyclotron.

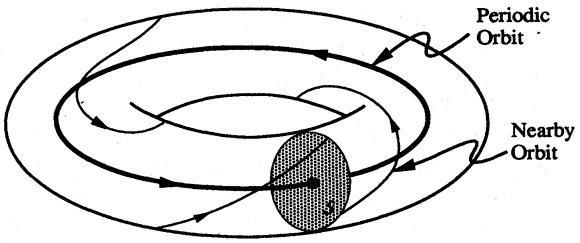

FIG. 8. Return mapping near an elliptic periodic orbit.

2. Poincaré section

Consider a two-degree-of-freedom system $H(p,q)$ with a periodic orbit. We construct a Poincaré section using a surface transversely intersecting this orbit at some point (Fig. 8). The mapping T from $\mathcal{S}$ to itself is defined locally near the periodic orbit, because points near the periodic orbit must return to $\mathcal{S}$, by continuity. The periodic orbit becomes a fixed point of T.

Suppose that the periodic orbit is elliptic. By definition, an orbit is elliptic if the return map T has a linearization M with eigenvalues $e^{\pm 2\pi i\omega}$ (see Sec. II.C). When ω is irrational there is a formal perturbation expansion for the map in terms of polarlike coordinates (r,θ) near the fixed point (Arnol'd, 1978, Appendix 7; Arrowsmith and Place, 1990, Chapter 6). In these coordinates the map is said to be in Birkhoff normal form:

$$T:\begin{cases} r'=r+h(r,\theta)\ , \\ \theta'=\theta+2\pi\omega+\rho_2 r^2+\cdots+\rho_{2m}r^{2m}+g(r,\theta)\ , \end{cases} \tag{1.37}$$

where h and g are $o(r^{2m})$, and m can be made as large as one likes. The map preserves the area $r\,dr\,d\theta$. If any of the ρ_{2n} are not zero, then the map has twist, providing r is small enough (the twist is to the right or to the left depending upon the sign of the first nonzero ρ). If we neglect h and g, then the radial coordinate is a constant, while θ rotates with a frequency depending on r—this is in fact the meaning of the twist condition. Typically this frequency is irrational, so the orbits tend to fill out the circles $r=$constant. Extending this to the full phase space, we see that the orbit lies on a torus. To the extent we can neglect g and h, there is a family of nested tori. However, the formal power series (1.37) does not converge in general, and some of the nested tori do not exist (see the discussion of the KAM theorem in Sec. III.B).

3. Incommensurate states

Symplectic maps also arise in condensed-matter physics. The simplest model of interest is a one-dimensional chain of particles connected by harmonic springs (Fig. 9). For simplicity, we take the spring constants to be 1. We can imagine this chain to be deposited on the surface of a crystal, which is represented by a periodic potential $V(x)=k/4\pi^2\cos(2\pi x)$. The conflict between the potential and the interatomic forces can result in an equilibrium state if force balance is satisfied:

$$(x_{j+1}-x_j)-(x_j-x_{j-1})+\frac{k}{2\pi}\sin(2\pi x_j)=0\ . \tag{1.38}$$

If we define $y_j=x_j-x_{j-1}$, and reinterpret the particle index j as "time," then this becomes the standard map (1.36). This model is known as the Frenkel-Kontorova model (Aubry, 1983b). The energy of a configuration is

$$W=\sum_j \tfrac{1}{2}(x_j-x_{j+1})^2+\frac{k}{4\pi^2}\cos(2\pi x_j)\ . \tag{1.39}$$

We shall learn much about this function and its extrema in Secs. V–VII.

4. Convex billiards

Consider a particle bouncing with elastic reflections in a bounded, two-dimensional domain (Berry, 1981). Since energy is conserved, the motion is completely determined by the sequence of boundary points at which the bounces occur. If the domain is convex, then the map from one bounce to the next is continuous. Convenient coordinates are *Birkhoff coordinates* (s,θ) (Fig. 10). The bounce position is measured by the arc length s along the boundary from a given point. The direction of motion is mea-

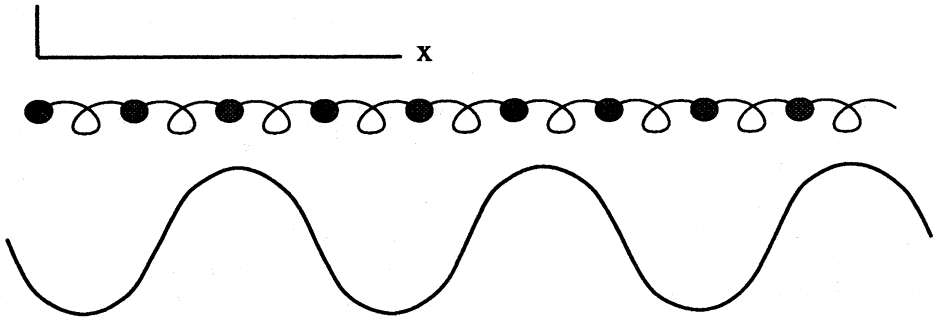

FIG. 9. Frenkel-Kontorova model.

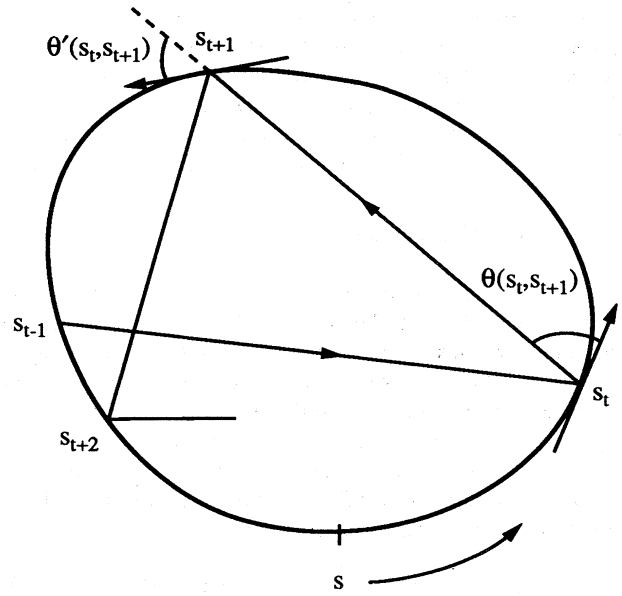

FIG. 10. Birkhoff coordinates for a billiard.

sured by the angle θ between a tangent to the boundary and the trajectory. It is easy to see that $s'(s,\theta)$ is a monotone increasing function of θ because of the convexity of the boundary (Fig. 11)—thus the map in Birkhoff coordinates has twist.

In fact, s is an anglelike coordinate since the map is periodic with period equal to the length of the boundary. As we shall see in Sec. V.C, this map preserves the area element $\sin\theta\, ds\, d\theta$. Thus canonical coordinates are given by $(x,y)=(s,\cos\theta)$. We could have anticipated this, since y is proportional to the component of the velocity along the boundary and is therefore the canonical conjugate of the arc length. The boundaries $y=\pm 1$ are fixed points and the twist, dx'/dy, vanishes at these points.

II. PHENOMENOLOGY

In this section we discuss range of phenomena that occur in twist maps. We use the standard map (1.36) as an illustration. It is in many ways a typical example of a smooth one-parameter family of area-preserving twist maps; however, it has two special aspects, which we discuss briefly.

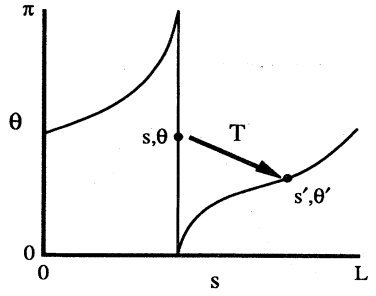

FIG. 11. Twist condition for billiard map in Birkhoff coordinates.

First, the standard map is special because it is periodic in the momentum direction—if $(x,y)\to(x',y')$, then the point $(x,y+m)\to(x'+m,\, y'+m)$, which is equivalent to $(x',y'+m)$ on the cylinder. So the orbits of two points separated by an integer in y are identical. We can use this to restrict our attention to the interval $0\le y\le 1$.

Second, the standard map is *reversible*—it has a time-reversal symmetry (DeVogelaere, 1950; Devaney, 1976; Sevryuk, 1986). Simple examples of reversible systems include Hamiltonians even in the momentum $H(-\mathbf{p},\mathbf{q})=H(\mathbf{p},\mathbf{q})$. In this case the time reverse of an orbit can be obtained by reversing the momentum. There is a similar time-reversal operator for the standard map. Reversibility is often used to help find periodic orbits (Greene, 1979); however, we do not discuss it further.

The sections are organized by increasing levels of chaotic behavior. We begin at $k=0$.

A. Integrable case

When $k=0$, the standard map becomes

$$\begin{aligned} y' &= y \ , \\ x' &= x + y' \ . \end{aligned} \tag{2.1}$$

Thus y is a constant of the motion, and x grows at a constant rate, which, however, increases with y because of the twist condition. Since the solution can be obtained in closed form,

$$y_t = y_0, \quad x_t = x_0 + y_0 t \ , \tag{2.2}$$

the map is "integrable."

1. Liouville integrability

In general, a symplectic map is *integrable* when the motion is "simple" in some way. To avoid philosophical issues (Zakharov, 1991) we shall consider only the notion of integrability in the sense of Liouville.

An *integral* is a function on the $2N$-dimensional phase space $I(z)$, which is invariant under the map:

$$I(T(z))=I(z) \ . \tag{2.3}$$

We wish to exclude the constant function, which is trivially invariant, so we assume

$$\nabla I \neq 0 \tag{2.4}$$

everywhere. This implies that $I=$constant defines a $(2N-1)$-dimensional surface or set of surfaces in the phase space. Assume that these are compact.

A set of n integrals $\{I^1, I^2, \ldots, I^N\}$ is *independent* if their gradients span an N-dimensional vector space at each point in phase space. Furthermore, the set is in *in-*

volution if all the mutual Poisson brackets vanish:

$$\{I^j, I^k\} = 0 \ . \tag{2.5}$$

Using these ingredients, we can state the Arnol'd-Liouville theorem for maps,

Theorem. *If there are N independent integrals in involution, then the motion lies on a nested family of N-dimensional tori, and there exist angle coordinates θ such that the map can be written in the form*

$$\begin{aligned} \mathbf{I}' &= \mathbf{I} \ , \\ \boldsymbol{\theta}' &= \boldsymbol{\theta} + \boldsymbol{\Omega}(\mathbf{I}) \ . \end{aligned} \tag{2.6}$$

Sketch of proof. Let M_c be a connected component of the set $\{z: I^i(z) = c^i, \ i = 1, \ldots, N\}$. Arnol'd has shown that if M_c is compact and connected, then it must be an N-torus (Arnol'd, 1978, Chapter 10). A construction of Darboux shows that, given a set of N independent functions I^j in involution, one can locally obtain canonically conjugate variables, θ^j, that is, $\{\theta^i, I^k\} = \delta_{jk}$. Because M_c is a torus, the θ^j can be chosen as angle variables. Since the Poisson bracket is preserved by a symplectic map, we have

$$\{\theta'^j, I'^k\} = \{\theta^j, I^k\} \ .$$

Using the invariance of I^k we can write this as

$$\{\theta'^j - \theta^j, I^k\} = 0 \ .$$

Since this is true for each j and k, the difference $\boldsymbol{\theta}' - \boldsymbol{\theta}$ can be a function only of the integrals. This function is $\boldsymbol{\Omega}(\mathbf{I})$ in (2.6).■

Thus the standard map, with $k = 0$, has a form that is typical of the integrable case, except that the frequency is linear in the momentum.

There are many examples of integrable maps (McMillan, 1971; Veselov, 1988; Quispel *et al.*, 1989; Bruschi *et al.*, 1991), though many of them have singularities in the phase space. In fact, the time t map of any one-degree-of-freedom, time-independent Hamiltonian is integrable. From our perspective, particularly interesting examples were found by Suris (1989), who showed that a map of the standard form (1.36) has a holomorphic (i.e., analytic in some domain) integral of the form

$$I(x,y) = F(x,y) + kG(x,y)$$

only when the sin( ) function is replaced by one of three forms, each of which has a number of parameters. One of these is periodic in x, and a special case is

$$\begin{aligned} y' &= y - \frac{1}{\pi} \arctan\left[\frac{k \sin(2\pi x)}{2 + k\cos(2\pi x)}\right] , \\ x' &= x + y' \ . \end{aligned} \tag{2.7}$$

This map has the integral

$$I(x,y) = \cos 2\pi y + k\{\cos(2\pi x) + \cos[2\pi(x-y)]\} \ . \tag{2.8}$$

For $k > 0$ there is an elliptic fixed point at (0,0) and a hyperbolic point at $(\frac{1}{2}, 0)$. There is also a pair of period-2 orbits; the orbit beginning at $(0, \frac{1}{2})$ is hyperbolic, and the other, at $\cos(2\pi x) = -k/2, \cos(2\pi y) = k^2/2 - 1$, is elliptic. Since $\nabla I = 0$ at these points, Eq. (2.7) is not strictly speaking Liouville integrable; however, all other invariant curves are topologically circles. For k small, (2.7) approaches the standard map.

2. Frequency

Since the standard map is defined on the cylinder, x should be taken mod 1 in Eq. (2.1). Thus y determines the rate of rotation around the cylinder.

In general, to define the rotation rate, we "lift" the angle coordinates to the real line. For the standard map this corresponds to computing $x' = x + y'$ without taking the fractional part. The *frequency* is defined as the limit

$$\omega = \lim_{t \to \infty} \frac{x_t}{t} \ , \tag{2.9}$$

if it exists. For Eq. (2.2) we have, trivially, $\omega = y$, for any initial condition.

The lift is not unique, because we could also use the equation

$$x' = x + y' + m$$

for any integer m to compute x'. Should we choose $m \neq 0$, the frequency would shift by m; we fix the lift by choosing $m = 0$.

3. Periodic and quasiperiodic orbits

There is an important distinction between rational and irrational values of ω. For each rational ω, the orbits of (2.1) are periodic on the cylinder. Generally an orbit is *periodic* with period n if n is the smallest integer such that

$$\begin{aligned} y_n &= y_0 \ , \\ x_n &= x_0 + m \end{aligned} \tag{2.10}$$

for some integer m. We shall denote such an orbit by (m,n). For an (m,n) orbit, the frequency always exists and is given by m/n.

Because of the twist condition, rational values of ω occur at a dense set of values of y; for the integrable case, these are just the values $y = m/n$.

On the other hand, almost all points have irrational ω. When ω is irrational the orbit never returns to its initial condition. An orbit is *quasiperiodic* if the frequency is irrational and the orbit is *recurrent*: it returns arbitrarily close to its initial condition. For the integrable map, when y is irrational, the x coordinate densely covers the circle $y = $constant on the cylinder. Thus these orbits are quasiperiodic.

The phase space of the integrable map is thus foliated by *rotational invariant circles.* A circle is "rotational" if

it encircles the cylinder (i.e., it is topologically equivalent to the circle $y=0$). For rational y the circles consist of an infinite number of periodic orbits; for irrational y they consist of an infinite number of quasiperiodic orbits, each of which is dense.

B. Nearly integrable case

As k is increased from zero, how much of the structure of the integrable map persists? A computer experiment, such as that shown in Fig. 12, can give some indication. One sees that most orbits still seem to lie on rotational curves; these orbits march around the cylinder in an ordered fashion and densely cover a circle. They seem to imply the existence of an integral $I(x,y)=y-Y(x)$ and its corresponding conjugate angle variable. The Kolmogorov-Arnol'd-Moser (KAM) theorem, which will be discussed in Sec. III, does indeed predict that most of the invariant circles persist for k small.

1. Resonances

However, there are orbits that no longer lie on invariant circles. To see these we have to focus closely on the points of rational frequency. For example, near $y=0$, where there was a circle of points of frequency 0/1, Fig. 12 shows an "island." This consists of a family of curves that are circles, but that do not encircle the cylinder. These are librational, as opposed to rotational, circles. The island is bounded by a *separatrix*, which is a curve separating the librational and rotational circles. We call the region of phase space bounded by the separatrix a "resonance zone" or simply a *resonance.*

This structure is analogous to the phase space of the pendulum. In fact, it is easy to see, if we consider the appropriate limit of the standard map, that the pendulum structure arises. Taking k and y both small implies that the differences in Eq. (1.36) can be replaced by derivatives

$$T\xrightarrow[k,y\to 0]{}\begin{cases}\dot{y}=-\dfrac{k}{2\pi}\sin(2\pi x)\ ,\\ \dot{x}=y\ .\end{cases} \tag{2.11}$$

Equations (2.11) are the differential equations for the pendulum. They have an invariant that is also the Hamiltonian for the system:

$$H=\frac{1}{2}y^2-\frac{k}{4\pi^2}\cos(2\pi x)\ . \tag{2.12}$$

There are two fixed points for the pendulum, (0,0) and $(\frac{1}{2},0)$, corresponding to the pendulum at rest either in the stable, downward position or in the unstable, upward position. The contours of $H<k/4\pi^2$ are librational circles. $H=k/4\pi^2$ is the separatrix, and $H>k/4\pi^2$ are rotational circles. Thus the pendulum has an island around $y=0$. This island has a full-width, its maximal extent in y, of

$$W=\frac{2}{\pi}\sqrt{k}\ . \tag{2.13}$$

Corresponding to this structure, the standard map also has only two fixed points (i.e., period-1 orbits) for $k\neq 0$, also at the points (0,0) and $(\frac{1}{2},0)$, see Fig. 12. When k is small, the size of the island also grows in accord with Eq. (2.13) as $\sqrt{k}$. As k increases, the shape of the standard map island begins to distort (it is not reflection symmetric about y axis) and its width grows more slowly

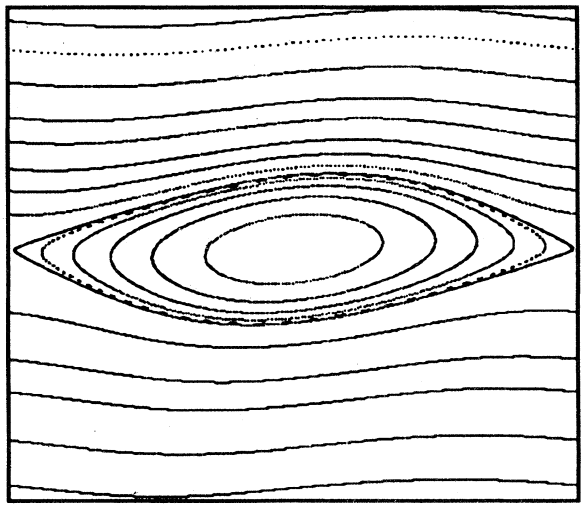

FIG. 12. Standard map phase space for $k=0.2$. Bounds are $[-0.5,0.5]$ on both x and y. Shown are many rotational invariant circles and the (0,1) resonance.

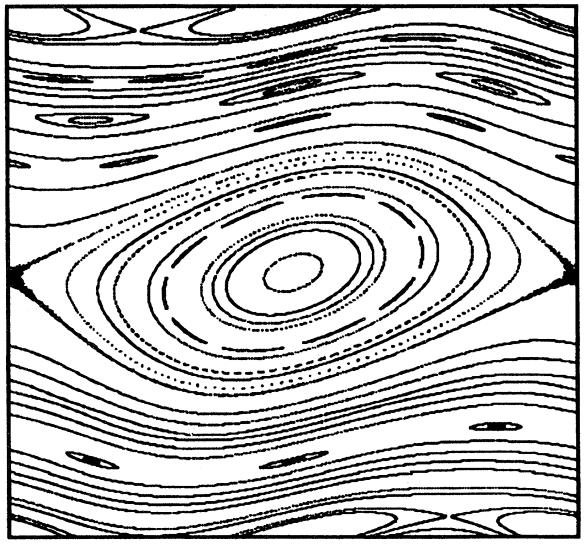

FIG. 13. Standard map for $k=0.5$. The (1,2), (2,5), (1,3), (1,4), (0,1), $(-1,3)$, and $(-1,2)$ resonances are shown. The separatrix of the (0,1) resonance exhibits a small amount of "fuzziness," i.e., "chaos."

than predicted by the pendulum approximation.

There are also resonances for other rational values of y, corresponding to the periodic orbits with frequencies m/n. Each resonance consists of a chain of n islands, and each island has a structure similar to the pendulum; several of these are shown in Fig. 13. Perturbation theory implies that the width of the m/n resonance grows as $k^{n/2}$ for k small. At the center of the island, and at the cusp of the separatrix, are periodic orbits with frequency m/n; typically there appear to be only two such periodic orbits. The existence of at least two orbits follows from the Poincaré-Birkhoff theorem, which we shall discuss in Sec. VI.

Orbits trapped in an island move successively from one island to another, following the periodic orbit (they skip $m-1$ islands each step). Thus there is an entire region of phase space that has frequency m/n.

2. Stability

To understand the structure of the orbits in the neighborhood of the periodic orbits, we consider their linear stability. Points in the neighborhood of an orbit (2.10) evolve according to the tangent map (1.17). After n iterations, in the linear approximation,

$$\begin{aligned}\delta z_n &= \left[\frac{d}{dz_0}T(T(\cdots T(z_0)))\right]\delta z_0 \\ &= M(z_{n-1})M(z_{n-2})\cdots M(z_0)\delta z_0 \\ &\equiv M^n \delta z_0 \ . \end{aligned} \tag{2.14}$$

Here $M(z)$ is the Jacobian matrix given by the derivative of $T(z)$. Since M is symplectic, so is M^n; and (1.23) implies that if λ is an eigenvalue, then so is $1/\lambda$. Here λ is a solution of the characteristic polynomial $\lambda^2-\mathrm{Tr}(M^n)+1=0$:

$$\lambda=\tfrac{1}{2}\{\mathrm{Tr}(M^n)+\sqrt{[\mathrm{Tr}(M^n)]^2-4}\} \ . \tag{2.15}$$

The possible stability properties are

(a) *hyperbolic*: both eigenvalues are real and larger than 1;

(b) *elliptic*: there is a pair of complex conjugate eigenvalues with unit modulus;

(c) *reflection hyperbolic*: both eigenvalues are real and less than 1;

(d) *parabolic*: the eigenvalues are both 1 or both -1.

These are summarized in Table I.

A stability classification is most conveniently given in terms of the *residue* (Greene, 1979);

$$R=\tfrac{1}{4}[2-\mathrm{Tr}(M^n)] \ . \tag{2.16}$$

The elliptic case, corresponding to $\lambda=e^{2\pi i\omega}$ or $0<R=\sin^2(\pi\omega)<1$, is the only one that could possibly be called stable, although the stability is a neutral one.

TABLE I. Stability classification.

Stability	λ	R	$\mathrm{Tr}(M)$
hyperbolic	>0	<0	>2
elliptic	$e^{2\pi i\omega}$	(0,1)	$(-2,2)$
reflection hyperbolic	<0	>1	<-2

We have already mentioned that near an elliptic periodic orbit with ω irrational, the mapping has a formal series representation (1.37), which has librational invariant circles. The full apparatus of the KAM theorem can be used to show that the orbit is generically stable (that is, points initially close stay nearby; Arnol'd, 1978), providing $\omega\neq m/n$ with $n<4$.

Positive residue corresponds to either an elliptic or a reflection hyperbolic orbit. These two cases are properly thought of as two manifestations of the same orbit. Negative residue always corresponds to a hyperbolic orbit. Finally, the parabolic case, $R=1$ or $R=0$, corresponds to points of bifurcation, where an orbit can cease to exist or lose stability.

For the standard map, the matrix M is

$$M=\begin{bmatrix}1-k\cos(2\pi x) & 1\\ -k\cos(2\pi x) & 1\end{bmatrix} , \tag{2.17}$$

which has the residue

$$R=\frac{k}{4}\cos(2\pi x) \ . \tag{2.18}$$

Thus the fixed point (0,0) has positive residue for $k>0$. It is elliptic for $0<k<4$ and becomes reflection hyperbolic for $k>4$. The point $(\frac{1}{2},0)$ is hyperbolic for $k>0$.

3. Stable manifolds

For a hyperbolic period-n orbit, M^n has two eigenvectors corresponding to the unstable and stable directions ($\lambda_1>1$ and $\lambda_2=1/\lambda_1<1$, respectively). Under M^n, points move on the branches of a hyperbola, with these eigenvectors as asymptotes. The stable manifold theorem (Lanford, 1973) implies that the eigenvectors of M^n can be extended to invariant manifolds W^u and W^s of T^n (Fig. 14). Each point on these accumulates on the hyperbolic orbit in at least one direction of time:

$$\begin{aligned} &z\in W^s \Longrightarrow T^{jn}z\to z_0 \ \text{ as } j\to\infty \ , \\ &z\in W^u \Longrightarrow T^{jn}z\to z_0 \ \text{ as } j\to-\infty \ , \end{aligned} \tag{2.19}$$

where z_0 is some point on the orbit. It is important to remember that the manifolds, while having the appearance of trajectories of a flow, are collections of orbits. A point on W^u, for example, moves a discrete distance upon application of T^n to another point on W^u. The stable manifold cannot intersect itself or the stable manifold of any other periodic orbit, since this would violate

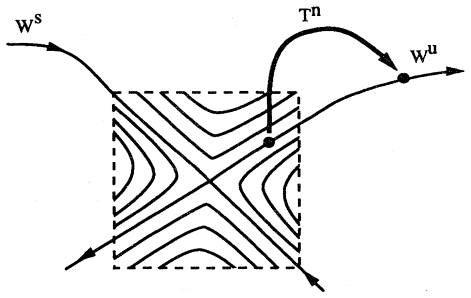

FIG. 14. Stable and unstable manifolds for hyperbolic fixed point. The hyperbolas with asymptotes given by the eigenvectors of the orbit, shown inside the box, give the local behavior. A global extension of the stable and unstable eigenvectors yields the stable and unstable manifolds.

uniqueness. Generically W^u and W^s are different manifolds; one exception to this is an integrable system for which W^u and W^s join smoothly to form a separatrix.

When W^u and W^s intersect transversely, the intersections are called *homoclinic* points. A homoclinic point lies on both the stable and the unstable manifold; so it is asymptotic to the hyperbolic orbit in both directions of time. Thus each iterate of a homoclinic point is also homoclinic, and the set of such iterates is a homoclinic orbit. *Heteroclinic* points are the intersection points of the stable and unstable manifolds of different periodic orbits.

Let z be a homoclinic point, as shown in Fig. 15. In addition to z, we shall see that there must also be a second homoclinic point ζ on W^u between z and its iterate $T(z)$. Let $\mathcal{A}$ be the closed region bounded by the curves W^u from z_0 to z and W^s from z to z_0. Since z is on W^s, $T(z)$ must be on the segment of W^s between z and z_0; however, at the next crossing along this segment, W^u must enter $\mathcal{A}$ (Fig. 16). This cannot occur at $T(z)$, since then orientation would be reversed. We label this crossing ζ; its orbit is homoclinic and distinct from that of z.

In fact, if there is one homoclinic orbit, there are an infinity of them (Poincaré, 1892). For example, some iterate of the segment of W^u between ζ and $T(z)$ must cross W^s. Consider the lobe $\mathcal{L}$ formed by the segments

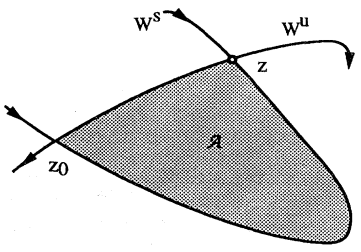

FIG. 15. Homoclinic intersection of the stable and unstable manifolds of z_0 at the point z.

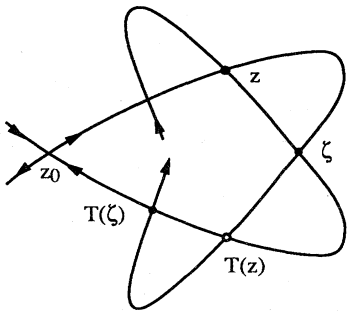

FIG. 16. Existence of a second homoclinic orbit defined by z.

of W^u and W^s between ζ and $T(z)$. By definition $\mathcal{L}$ is contained in the region $\mathcal{A}$; if the segment of W^u were never to cross W^s, then all future iterates of $\mathcal{L}$ must remain in $\mathcal{A}$. However, the area of $\mathcal{L}$ is preserved under iteration, and since the area of $\mathcal{A}$ is finite, $T^j(\mathcal{L})$ cannot remain in $\mathcal{A}$ forever. Thus this segment of W^u must eventually cross W^s, giving rise to at least two new homoclinic points (see Fig. 17). We shall discuss this process in Secs. VIII and IX.

C. Transition

1. Destruction of invariant circles

As k increases, the resonances grow in size, and the region of phase space occupied by rotational invariant circles necessarily shrinks. Invariant circles are destroyed when resonances engulf the region of phase space they once occupied. In Fig. 18 we show the standard map at a moderate parameter value for which most of the invariant circles are destroyed.

The twist condition implies that the frequency is essentially a monotonic function of y (this will be made precise in Sec. VI). Thus a rotational invariant circle of frequency ω must lie between any pair of resonances whose frequencies surround it. Suppose there is a heteroclinic connection between this pair; that is, they *overlap*. There can then be no such invariant circle. The heteroclinic

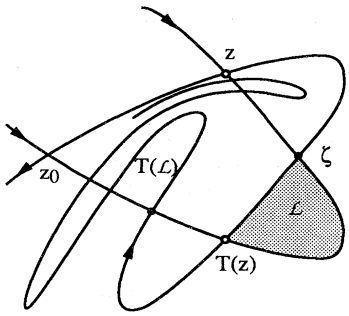

FIG. 17. Existence of infinitely many homoclinic intersections.

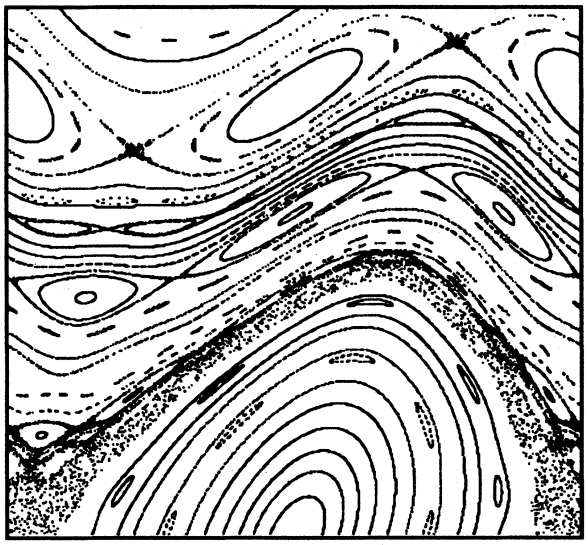

FIG. 18. Standard map for $k=0.8$. Bounds are $x\in[-0.5,0.5], y\in[0.0,0.6]$. There are still some visible rotational invariant circles.

connection implies that there is a curve formed from a segment of unstable manifold of one resonance and stable manifold of the other, which crosses the region that was to have contained the invariant circle. Every point on this curve is either forward asymptotic to one resonance or backward asymptotic to the other and thus cannot be on an invariant circle.

Chirikov has introduced a perturbative technique for computing the overlap of resonances and therefore the parameter for the destruction of invariant circles (Chirikov, 1979b). The method is to approximate the map by the pendulum Hamiltonian in the neighborhood of the resonance and to use this to estimate the resonance widths. For example, aside from the (0,1) resonance, the standard map also has a (1,1) resonance corresponding to the elliptic point (0,1) and the hyperbolic point $(\frac{1}{2},1)$. Setting $y=1+\delta y$ and $x=t+\delta x$ in Eq. (1.36), and approximating for k and δy small, we obtain the pendulum equations just as in (2.11). Thus the width of the (1,1) resonance $W_{(1,1)}$ is equal to $W_{(0,1)}$, as given by Eq. (2.13). The distance between these resonances is 1; and so, as shown in Fig. 19, they overlap when

$$\frac{1}{2}(W_{(0,1)}+W_{(1,1)})=1 \Longrightarrow k=\frac{\pi^2}{4}\ . \tag{2.20}$$

This rough estimate would predict that there are no invariant circles in the range $0<\omega<1$ when $k>2.5$. In fact, this is a considerable overestimate of the actual overlap value, since the pendulum approximation is valid only for small k. Considerable improvement can be obtained by higher-order perturbation theory (Chirikov, 1979b; Lichtenberg and Lieberman, 1982).

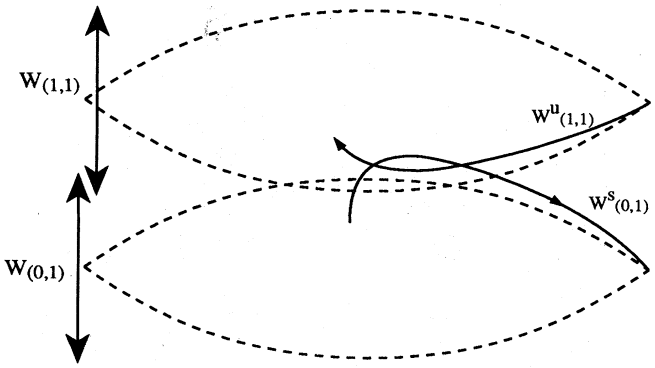

FIG. 19. Resonance overlap. Pendulum approximation gives the dashed shapes for the (0,1) and (1,1) resonances. Actual stable and unstable manifolds are sketched as the solid curves. Intersection points of these are heteroclinic orbits.

2. Last invariant circle

As k increases, there are fewer invariant circles. In fact, as we shall see in Sec. IV, there is a simple analytic argument that shows when k is large enough there can be no invariant circles (the "converse KAM" theorem). A natural question to ask is, which invariant circle is the last?

Greene (1979) discovered that for the standard map, the last invariant circle has frequency $\omega=\gamma$, where γ is the golden mean

$$\gamma=\frac{1+\sqrt{5}}{2} \tag{2.21}$$

(special symmetries of the standard map imply that all the circles with frequencies $m\pm\gamma$ are destroyed simultaneously; we refer to this set of circles as the "golden circle"). Greene developed a method for determining the existence of an invariant circle by looking at the stability of nearby periodic orbits. He reasoned that if there is a set of periodic orbits whose frequencies limit on the invariant circle.

$$\lim_{i\to\infty}\frac{m_i}{n_i}\to\omega \tag{2.22}$$

and which have residues between zero and 1, then the invariant circle will exist. This "residue conjecture," has been proved in some cases (MacKay, 1991).

A natural set of frequencies to use is that given by the continued-fraction convergents of ω (we shall discuss these in Sec. III). In this case the parameter values for which the ith convergent has $R=1$ geometrically limits on a value $k_{\rm cr}(\omega)$, which is the parameter at which the invariant circle is destroyed. For the golden mean this value is

$$k_{\rm cr}(\gamma)\approx 0.971\,635\,406\ . \tag{2.23}$$

We show the standard map phase space at $k_{\rm cr}(\gamma)$ in Fig.

FIG. 20. Standard map for $k=0.971\,635\,4$. The invariant circle shown between the (1,3) and (2,5) resonances has frequency $1/\gamma^2$—it is equivalent by symmetry to the golden circle. Close examination fails to reveal any other rotational invariant circles.

20. There is a remarkable self-similarity associated with this parameter value, and the application of renormalization-group ideas has been very fruitful (MacKay, 1983). Since these methods have been reviewed elsewhere, we shall not discuss them further (MacKay, 1986).

Using the Greene method one can construct a "fractal diagram" of parameter values $k_{\rm cr}(m/n)$ such that $R=1$ for $\omega=m/n$ (Schmidt and Bialek, 1982). On this diagram the golden circle has the largest $k_{\rm cr}$. An alternative method, estimating the radius of convergence of a Fourier series for the invariant circle, leads to a similar conclusion (Percival, 1982).

3. Islands around islands

The entire structure we have just discussed is also found in the neighborhood of any elliptic periodic orbit.

An elliptic period-n orbit is a fixed point of the map T^n. The linearization about this point has orbits rotating with the frequency ω (recall Table I). Near the fixed point the map can be expanded and written in the form (1.37); if any of the ρ_{2k} are nonzero, the map has twist in some neighborhood of the point. Thus nearby orbits rotate about the fixed point, and the rotation frequencies vary with the distance away from the fixed point.

Thus as one moves away from the fixed point the rotation frequency must go through rational values, and at each such point a resonance is formed. If the rational number is m_1/n_1, then the resonance corresponds to a fixed point of $(T^n)^{n_1}$. We call these orbits of *class 1* (rotational orbits have class zero). In the neighborhood of any class-1 elliptic periodic orbit the same structure repeats. Thus we expect to see the structure of islands around islands, and it can even occur in a self-similar way (Meiss, 1986; see also Fig. 21). This structure was already envisioned by Birkhoff (1935), who said,

> It is clear that not only do general elliptic periodic solutions possess neighboring elliptic and hyperbolic periodic solutions, but also, beginning again with the neighboring elliptic solutions, who are, as it were, satellites of these solutions, one can obtain other elliptic and hyperbolic solutions which are secondary satellites.

D. Chaos

As of yet we have not discussed the most intriguing phenomena that occur when k is increased, that of *chaos*. There are three basic ingredients for chaos (Devaney, 1986). First, one requires "sensitive dependence on initial conditions;" that is, nearby orbits should separate exponentially in time (positive Lyapunov exponents). Second, the motion should be bounded, so that the exponential separation does not simply result in smooth expansion to infinity. This means that separating orbits must eventually come close together again. Recurrent, but in practice unpredictable, behavior is a signature of chaos. Finally, there should be some large set of orbits (one of nonzero measure) that has this behavior.

A similar concept is Birkhoff's *irregular component*, a connected set that is the complement of the elliptic periodic orbits and invariant circles. Hyperbolic periodic orbits and their stable and unstable manifolds are part of an irregular component. In fact, their transversal intersection is a prime ingredient in chaos—giving rise to the famous Smale horseshoe structure (Moser, 1973). Existence of a horseshoe implies that there is a zero-measure set of orbits that act chaotically: they can be equivalent to a coin toss (Bernouilli shift). To our knowledge there are no results that imply a nonzero measure of orbits is chaotic for a typical system, and it is not known whether irregular components typically have nonzero measure. There are examples of completely ergodic systems, such as the Arnol'd cat map (which is a twist map; Arnol'd and Avez, 1968), and specially constructed examples of systems with both invariant circles and irregular components (Wojtkowski, 1981).

On the other hand, computer-generated pictures, e.g., Fig. 22, imply that the measure of a typical irregular component is nonzero; they seem to be "fat fractals" (Umberger and Farmer, 1985). Understanding the structure of these regions, and the way in which typical orbits move through them, is a major goal of the study of chaos.

1. Transport

The inherent loss of predictability for chaotic systems suggests that it is not especially efficient or useful to try

to follow individual trajectories. One alternative is to describe the properties of ensembles of trajectories. Thus even though we study deterministic systems, statistical methods may be appropriate.

Transport theory deals with the motion of ensembles of trajectories, asking how long it takes a set of orbits to move from one region of phase space to another. An understanding of transport properties allows one to compute transition probabilities and correlation functions.

Applications of transport include the calculation of chemical reaction rates. A chemical system can be modeled by a set of differential equations and the reaction itself by the transition between two regions of phase space. Such an approach was pioneered by Wigner (Wigner, 1937). Simple chemical reactions can be modeled by classical Hamiltonian systems, and an understanding of classical transport has been found to be useful (Davis, 1985; Skodje and Davis, 1988).

Another application is to mixing in fluids (Aref, 1984; Ottino, 1989; Rom-Kedar *et al.*, 1990). The motion of a passive scalar in a given velocity field can result in mixing, and the most efficient mixing occurs for chaotic velocity fields.

An understanding of transport is also important in plasma and accelerator physics. The basic problem here is to confine a set of interacting particles to one region of phase space, corresponding to the configuration being in the interior of the reactor and the momenta being large enough so that significant nuclear reactions can occur. For magnetic confinement of plasmas, the simplest model of such a system, guiding-center particle motion, can be reduced to an area-preserving map (Rosenbluth *et al.*,

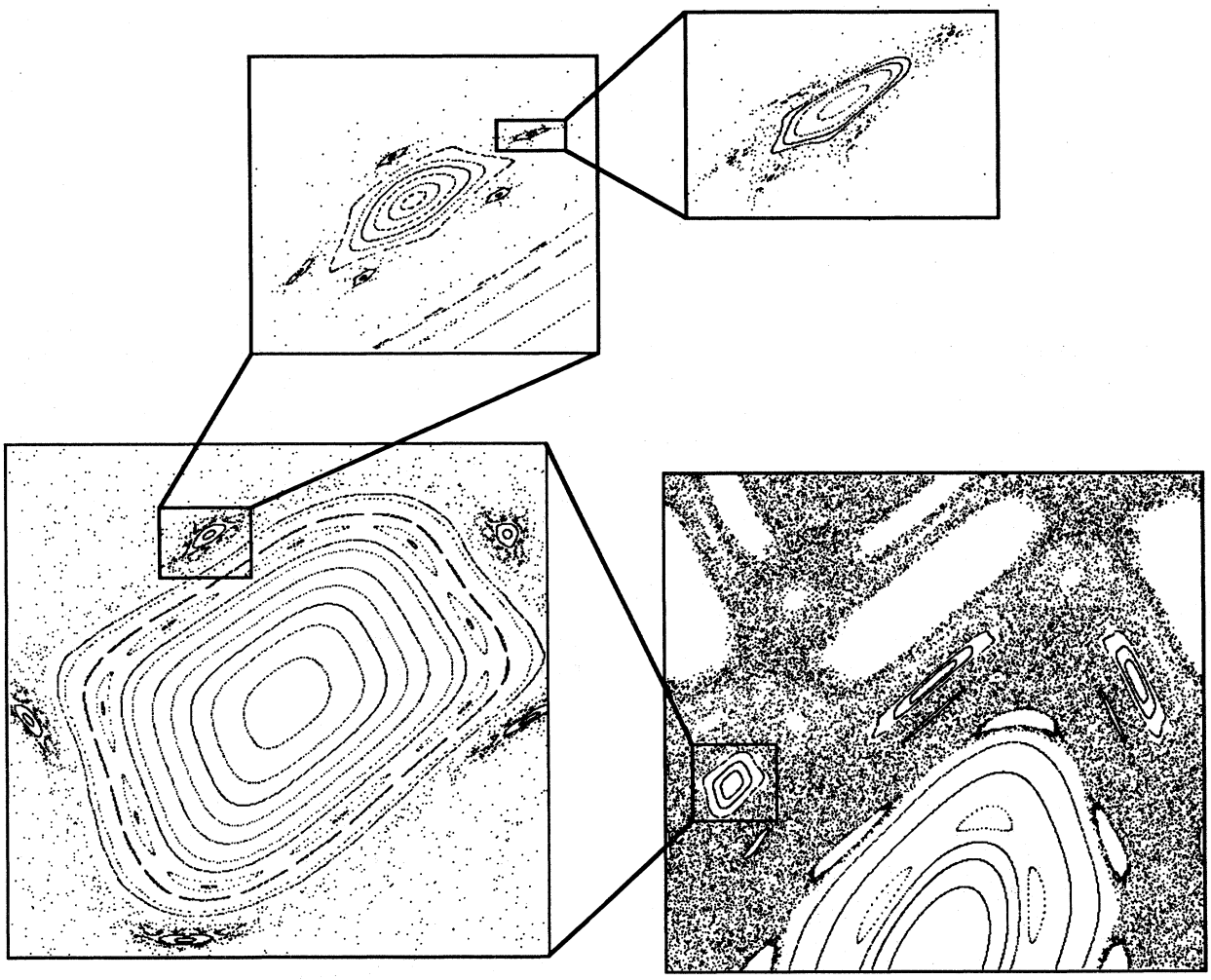

FIG. 21. Islands around islands for the standard map at $k = 1.201\,413\,33$. The bottom-right figure has the bounds $[-0.5, 0.5]$ for x and $[0.0, 0.6]$ for y. One island of the (1,3) resonance at the bottom right is enlarged in the figure to the left, revealing, among other things, a class-1 (1,5) island chain around it. This island, when enlarged, has a class 2 (1,5) chain, which when enlarged, etc. The parameter value was chosen to observe this self-similar structure (Meiss, 1986).

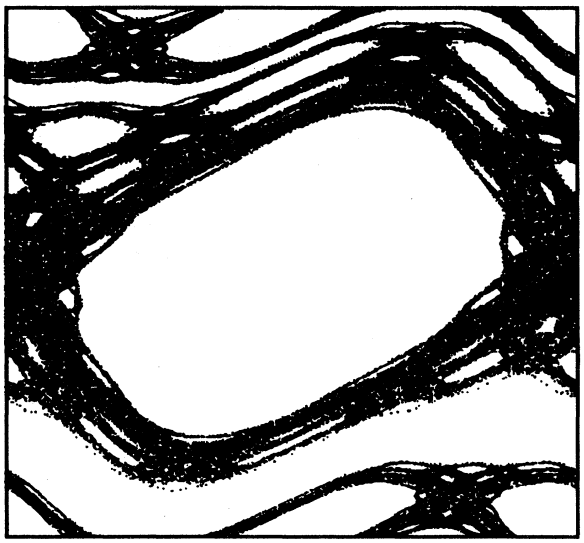

FIG. 22. Standard map for $k=1.0$. Shown are about 10^4 iterates of two chaotic orbits. These orbits appear to fill regions of nonzero area.

1966; Rechester and Rosenbluth, 1978). Accelerators are naturally modeled by maps (Carrigan *et al.*, 1982; Jowett *et al.*, 1986).

2. Flux

The most elementary transport problem is to determine the volume of trajectories that escape from some region per unit time, or *flux*. In the case of volume-preserving motion, the net flux is always zero; so one would like to compute the one-way flux.

Consider a map and a region bounded by a curve $\mathcal{C}$ (see Fig. 23). The flux $\mathcal{F}(\mathcal{C})$ is the area escaping from $\mathcal{C}$: the area inside $T\mathcal{C}$ that is also outside of $\mathcal{C}$. If $\mathcal{C}$ encloses finite area, then area preservation implies that the escaping flux is the same as the entering flux. If $\mathcal{C}$ is a rotational circle, the flux is the area above $\mathcal{C}$ that is below $T\mathcal{C}$. Of course when $\mathcal{C}$ is an invariant circle, it has zero flux.

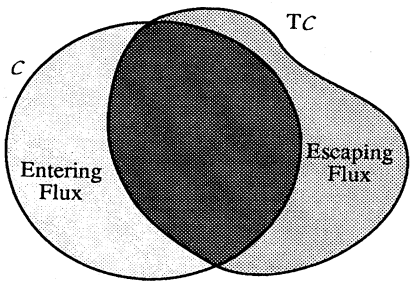

FIG. 23. Flux definition. The area outside $\mathcal{C}$ that is inside $T\mathcal{C}$ is the flux through $\mathcal{C}$.

Upon each iteration of the mapping, an area $\mathcal{F}$ escapes from $\mathcal{C}$ and the same amount enters. Thus the flux gives an estimate (sometimes a crude one) of a confinement time for $\mathcal{C}$: If motion in $\mathcal{C}$ is "random" in some sense, then a trajectory will be trapped within $\mathcal{C}$ for a typical time

$$t_{\text{trapped}}=\frac{A(\mathcal{C})}{\mathcal{F}(\mathcal{C})}\ , \tag{2.24}$$

where $A(\mathcal{C})$ is the area enclosed by $\mathcal{C}$. A better estimate of confinement time for the irregular trajectories would be obtained if A were replaced by the area of the connected irregular component inside $\mathcal{C}$. However, this is difficult to determine.

As an example, we compute the flux across the circle $\mathcal{C}=\{y=y_0\}$ for the standard map. The iterate of $\mathcal{C}$, by Eq. (1.36), is the curve

$$T\mathcal{C}=\{y=y_0-k\sin[2\pi(x-y_0)]/2\pi\}\ . \tag{2.25}$$

The upward flux is the area above $\mathcal{C}$ and below $T\mathcal{C}$, which is

$$\mathcal{F}=k/2\pi^2\ . \tag{2.26}$$

This is also the downward flux.

3. Diffusion

When k is large, say of order 100, then the phase space of the standard map looks, to the resolution of a typical computer screen, completely chaotic. The rapid loss of phase coherence for large k makes it plausible that statistical approximations should be valid. Because the jump in y is proportional to $k\sin(2\pi x)$, an $O(\epsilon)$ uncertainty in x gives rise to an error $O(k\epsilon)$ in y. Using this in the x equation leads to an error $O(k\epsilon)$ in that of x. The exponential escalation in error, by a factor of order k each step, makes the phase completely undetermined after a small number of steps.

Since the step in y depends on the highly uncertain phase, x, we expect the motion of y to be diffusive in character. The diffusion coefficient is defined as the mean-square spread in y per step,

$$D\equiv\lim_{t\to\infty}\frac{\langle(y_t-y_0)^2\rangle}{2t}\ , \tag{2.27}$$

where the average $\langle\ \rangle$ can be thought of as an average over some ensemble of initial conditions. The factor of 2 in (2.27) appears in the Fokker-Planck derivation (Lichtenberg and Lieberman, 1982). Using $\Delta y_t=y_t-y_{t-1}$, we can write (2.27) as

$$\begin{aligned}D&=\lim_{t\to\infty}\frac{1}{2t}\sum_{i,j=1}^{t}\langle\Delta y_i\Delta y_j\rangle\ ,\\&=\lim_{t\to\infty}\frac{1}{2t}\sum_{i=1}^{t}\sum_{j=1-i}^{t-i}\langle\Delta y_i\Delta y_{i+j}\rangle\ .\end{aligned} \tag{2.28}$$

The average in (2.28) is the force correlation function

$$C_t \equiv \langle \Delta y_0 \Delta y_t \rangle \ . \tag{2.29}$$

Here we have noted that the average over initial conditions is time translation invariant, since the map is area preserving: $dx_0\, dy_0 = dx_j\, dy_j$. Reversing the order of the sums in (2.28) yields

$$D = \lim_{t\to\infty} \frac{1}{2} \sum_{j=1-t}^{t-1} \left[1 - \frac{j}{t}\right] C_j \ ,$$
$$= \frac{1}{2} \sum_{j=-\infty}^{\infty} C_j \ , \tag{2.30}$$

where the last sum is valid providing the correlations decay at least as rapidly as t^{-2}.

Formally, (2.30) can be applied to the standard map for any value of k. Whenever there are rotational invariant circles, then D must be zero, since, according to the definition (2.27), diffusion requires that the momentum reach arbitrarily large values. Thus for $k < k_{\rm cr}(\gamma)$ of (2.23), $D = 0$. We shall discuss the form of D for k slightly larger than $k_{\rm cr}(\gamma)$ in Sec. IX.

When k is large, the correlations should decay rapidly; the simplest statistical approximation is to assume that x is an uncorrelated random variable—the random-phase approximation. Then only C_0 is nonzero, and for the standard map $C_0 = k^2 \langle [\sin(2\pi x)]^2 \rangle / 4\pi^2 = k^2/8\pi^2$. Thus the diffusion becomes

$$D_{\rm QL} = \frac{k^2}{16\pi^2} \ , \tag{2.31}$$

which goes by the name "quasilinear diffusion." It is indeed observed that when k is large, D approaches $D_{\rm QL}$.

Corrections to $D_{\rm QL}$ can be systematically computed by including correlations in (2.30) for $j \neq 0$. This leads to a series in products of Bessel functions (Cary *et al.*, 1981; Rechester *et al.*, 1981). These results agree well with moderate time computations of the diffusion coefficient using either an ensemble of initial conditions or a single initial condition which is chosen to be in the chaotic region (Meiss *et al.*, 1983; Ichikawa *et al.*, 1987).

4. Long-time tails

However, the series for D does not appear to converge at many parameter values, because the assumption $C_t = O(t^{-2})$ in Eq. (2.30) fails.

The long-time behavior of correlation functions is a problem of continuing interest. Whenever there are regular regions, such as those caused by an elliptic periodic orbit, the correlation function appears to decay algebraically with time (Karney, 1983; Meiss *et al.*, 1983; Chirikov and Shepelyanksy, 1984; Geisel and Thomae, 1984). This occurs even when the average is taken over only chaotic orbits. The reason seems to be the stickiness of the regular orbits. Whenever a chaotic orbit wanders close to an invariant circle it stays close for a long time.

A related problem for the standard map is the existence of accelerator modes—orbits that satisfy $y_n = y_0 + j$, $x_n = x_0 + m$ for integers n, m, and j. These exist due to the periodicity in the y direction. Whenever there is an elliptic accelerator mode, the diffusion coefficient appears to be infinite (Karney *et al.*, 1982; Meiss *et al.*, 1983).

We shall discuss the long-time tail problem in Sec. IX.

III. NUMBER THEORY AND KOLMOGOROV-ARNOL'D-MOSER (KAM) THEORY

A. Number theory

The persistence of invariant circles for small perturbations from the integrable case depends on the fact that some irrational numbers are "far" from rationals. Here we discuss and quantify the degree of irrationality.

1. Diophantine numbers

An irrational number can be approximated arbitrarily closely by rational numbers whose denominators are arbitrarily large. However, some irrationals are more difficult to approximate than others. To measure this we use the distance $|n\omega - m|$ between a number and the rational m/n. We say ω is particularly hard to approximate if it satisfies a *Diophantine condition*: there exists a $C > 0$ such that for all integers $(m,n) \neq (0,0)$

$$|n\omega - m| > \frac{C}{n^\tau} \tag{3.1}$$

for some $\tau \geq 1$. Let $D_\tau(C)$ be the set of ω that satisfy (3.1). Equation (3.1) implies that ω is excluded from intervals surrounding each rational (Fig. 24). For C small enough, $D_\tau(C)$ is not empty; in fact, for any $\tau > 1$ the measure of $D_\tau(C)$ approaches 1 as C approaches zero (Khinchin, 1964). Consider, for example, the numbers in the interval (0,1]. The complement of $D_\tau(C)$ has a measure μ which is given by the sum of all the excluded intervals, each of which have a width $C/n^{\tau+1}$; thus

$$\mu(\bar{D}_\tau(C)) = \sum_{n=1}^{\infty} \sum_{\substack{m=1 \\ (m,n)\ \text{coprime}}}^{n} \frac{C}{n^{\tau+1}}$$
$$= C \sum_{n=1}^{\infty} \frac{\phi(n)}{n^{\tau+1}} = C \frac{\zeta(\tau)}{\zeta(\tau+1)} \ , \tag{3.2}$$

where $\phi(n)$, the Euler function, is the number of integers

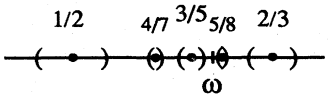

FIG. 24. Excluded intervals about rationals implied by Eq. (3.1).

not exceeding and relatively prime to n, and $\zeta(\tau)$, the Riemann zeta function, is finite when $\tau>1$ (Abramowitz and Stegun, 1965, Sec. 24.3.2). Thus $\mu(\overline{D}_\tau(C))\to 0$ as $C\to 0$ for any $\tau>1$.

The set of Diophantine numbers D_τ is the union of $D_\tau(C)$ for all $C>0$.

2. Continued fractions

Another classification of the properties of real numbers arises from continued-fraction expansions (Khinchin, 1964). The continued fraction of ω is the sequence $[a_0,a_1,\dots]$ of integers generated by the map

$$a_n=[\omega_n]\ ,\qquad \omega_{n+1}=\frac{1}{\omega_n-a_n}\ , \tag{3.3}$$

where the square brackets indicate the nearest integer less than ω (if ω is negative, a_0 is negative and the remaining a_i are positive), and $\omega_0=\omega$. An alternative representation for the continued fraction is

$$\omega=a_0+\cfrac{1}{a_1+\cfrac{1}{a_2+\cdots+\cfrac{1}{a_n+\cdots}}}\ . \tag{3.4}$$

The continued-fraction expansion of an irrational is infinite (since if ω_n is irrational, then ω_{n+1} is also irrational), while that for rationals always ends (one eventually finds that ω_{n+1} is an integer). Every rational has two equivalent continued-fraction representations:

$$[a_0,a_1,\dots,a_i]=[a_0,a_1,\dots,a_i-1,1]\ , \tag{3.5}$$

where $a_i\neq 1$ (unless $i=0$). *Convergents* of a continued fraction are the rationals obtained by truncating the expansion at some stage:

$$m_i/n_i=[a_0,a_1,\dots,a_i]\ , \tag{3.6}$$

where m_i and n_i are coprime. The continued-fraction expansion is a *strongly convergent* expansion: for any ϵ there is a j such that

$$|n_i\omega-m_i|<\epsilon \quad \text{for all } i\geq j\ .$$

Furthermore, the convergents are *best approximants*—if m/n is a convergent of ω, then every m'/n' with $n'\leq n$ is farther from ω: $|n'\omega-m'|<|n\omega-m|$.

Every convergent is close to the frequency that it approximates in the sense that it satisfies

$$|n\omega-m|<C/n \tag{3.7}$$

for $C=1$; conversely, every rational that satisfies (3.7) for $C=\frac{1}{2}$ is a convergent. However, when $C<1/\sqrt{5}$, there exist ω such that only finitely many convergents satisfy (3.7).

Irrationals are more difficult to approximate if their continued-fraction elements are small. This is because a large element a_{i+1} leads to a small correction to m_i/n_i. A prominent example of such behavior is the number π, which has the continued-fraction expansion

$$\pi=[3,7,15,1,292,1,1,1,2,1,3,1,14,2,1,1,2,2,2,1,84,2,\dots] \tag{3.8}$$

so that π is well approximated by its second convergent, 22/7, and its fourth convergent, 355/113. This leads to the definition of the numbers of *constant type*: those numbers for which there is an α such that $a_i<\alpha$ for all i. For such ω, and for sufficiently small C, there are no (m,n) satisfying the inequality (3.7). In fact, the numbers of constant type are precisely those that satisfy a Diophantine condition (3.1) for $\tau=1$. The set of numbers of constant type has measure zero.

A subset of the numbers of constant type are the *quadratic irrationals*: the solutions of a quadratic equation with integer coefficients. Lagrange showed that every quadratic irrational has an eventually periodic continued fraction, and conversely every eventually periodic continued fraction corresponds to a quadratic irrational. Quadratic irrationals are a special case of the *algebraic irrationals*: solutions of a polynomial of degree n with integer coefficients. Roth has shown that every algebraic irrational is in $D_{1+\delta}$ for any $\delta>0$ (Cassels, 1965).

A more special subset of the numbers of constant type are the *noble numbers*: these have $a_i=1$ for all i larger than some j. Noble numbers are dense in the reals, since one can append a noble tail to a convergent of any ω to obtain an arbitrarily good approximation to ω. On the other hand, the nobles are a set of measure zero, since they can be put in one-to-one correspondence with the rationals. The noblest of numbers is the golden mean (2.21),

$$\gamma=[1,1,1,\dots]\ . \tag{3.9}$$

Since (3.9) is periodic, γ is a quadratic irrational (in fact, it is the larger solution of $\gamma^2=\gamma+1$). Sometimes $1+\sqrt{2}=[2,2,2,\dots]$ is referred to as the *silver mean*; it is quadratic, but not noble.

We show the relation between these classifications in Fig. 25.

3. Farey tree

The Farey tree (Hardy and Wright, 1979) is a technique for organizing the rational numbers according to the length of their continued-fraction expansions. The tree is constructed beginning with a pair of rationals in

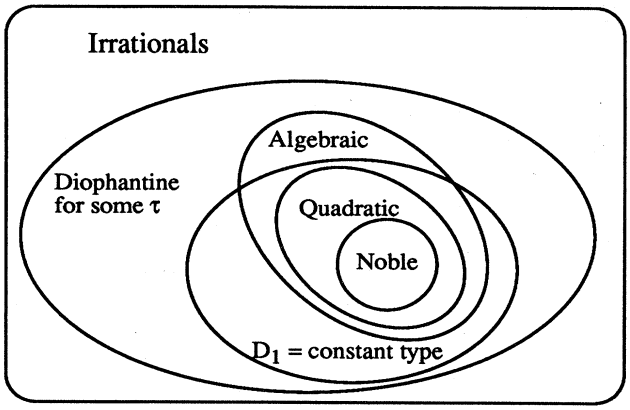

FIG. 25. Venn diagram of the irrational numbers.

lowest terms, m/n and m'/n', which are *neighboring*: $mn'-nm'=1$. Level one of the tree is generated from these by adding their numerators and denominators,

$$\frac{m''}{n''}=\frac{m+m'}{n+m'} . \tag{3.10}$$

This rational is the *mediant* of m/n and m'/n'. It is not difficult to see that m'' and n'' are coprime and that m''/n'' is a neighbor to both its parents. To construct the second level, find the mediants of m''/n'' and each of its parents. This construction leads to a binary tree that gives every rational number in the interval $[m'/n', m/n]$. The tree generated by the neighbors 1/0 and 0/1, shown in Fig. 26, gives all the positive numbers.

The Farey path for a number is the sequence of left/right steps leading to it from 1/1. Thus the Farey path for 2/7 is LLLR. Irrationals are represented by infinitely long Farey paths. The Farey path provides a binary code for the reals.

The continued-fraction expansion is closely related to the Farey tree construction. The sum of the continued-fraction elements of $m/n=[a_0,a_1,\ldots,a_i]$ gives the level on which it occurs:

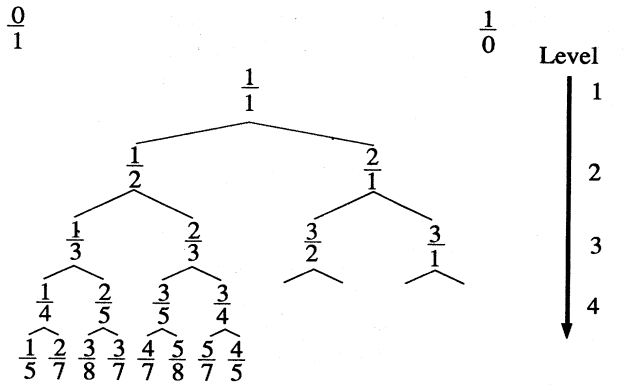

FIG. 26. Farey tree construction.

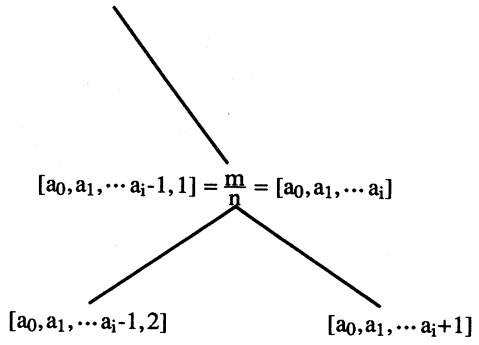

FIG. 27. Relation of the continued fraction to the Farey tree.

$$\text{Level}([a_0,a_1,\ldots,a_i])=\sum_{j=0}^{i} a_j . \tag{3.11}$$

The continued fraction for a given Farey path can be obtained recursively. The continued fraction for a daughter of m/n is obtained by incrementing a_i by 1. The two representations (3.5) give the two daughters; that with $a_i\neq 1$ is used if the current step in the Farey path is in the same direction as the preceding step, and that with $a_i=1$ if the direction changes (see Fig. 27). For example, the golden mean corresponds to the path RLRLR$\cdots=[1,1,1,1,1,1,\ldots]$. In general, noble numbers have a Farey path that eventually alternates, $\cdots$LRLR$\cdots$.

There are two different types of infinite Farey paths: those that eventually consist of all L's or all R's and those that continue to alternate. The former converge to rational numbers. For example, the sequence

$$\text{RLLLLLLLL}\cdots\rightarrow\frac{1}{1}\Big|_{+} \tag{3.12}$$

approaches 1/1 from above and

$$\text{LRRRRRRRR}\cdots\rightarrow\frac{1}{1}\Big|_{-} \tag{3.13}$$

from below. These two numbers should be thought of as distinct from 1/1—they have a nice interpretation in terms of the orbits of a twist map, as we shall see in Sec. VIII. Farey paths that never settle down to either one direction or the other approach irrational numbers.

B. KAM theory

Consider an integrable area-preserving map, Eq. (2.6), satisfying the twist condition (1.30). Thus

$$d\Omega/dI\geq K>0 . \tag{3.14}$$

The twist condition implies that there are quasiperiodic orbits for all irrational ω; in fact, since I is a constant of motion, the frequency is just $\Omega(I)$.

The KAM theorem, in this context, implies that rotational invariant circles with sufficiently irrational frequency persist under small area-preserving perturbations. A perturbation is small if it and its first j derivatives are small; to express this formally, define the j norm of a function f as

$$|f(x,y)|_j = \sup_{m+n\le j} \left| \frac{\partial^{m+n} f}{\partial x^m \partial y^n} \right| .$$

The perturbed map is written

$$\begin{aligned} I' &= I + f(I,\theta) , \\ \theta' &= \theta + \Omega(I) + g(I,\theta) . \end{aligned} \tag{3.15}$$

As we shall see in Sec. IV.B, in order that there be invariant circles it is necessary that the average of $f(I,\theta)$ be zero,

$$\int_0^1 f(I,\theta)\, d\theta = 0 , \tag{3.16}$$

since otherwise the perturbed map could simply shift all points vertically. For this case the KAM theorem is

Theorem (Moser, 1973). *If $\Omega(y)$ satisfies (3.14) and is j times differentiable, then there is an $\epsilon > 0$ such that all area-preserving maps (3.15) and (3.16) with $|f|_j + |g|_j < \epsilon K C^2$ have rotational invariant circles for all frequencies that satisfy a Diophantine condition (3.1) with*

$$1 < \tau < (j-1)/2 . \tag{3.17}$$

The theorem implies that the more differentiable a system is, the more invariant circles it has, since τ can be larger. The inequality (3.17) requires $j > 3$; however, Herman (1983, 1985) has shown that this theorem can be extended to the case $j = 3, \tau = 1$, providing an additional Hölder condition is imposed upon the third derivatives. Furthermore, he has given examples of perturbations which, being C^2 but not C^3 ($C^{3-\epsilon}$), do not have invariant circles.

One of the most important concepts arising from the KAM theorem is the labeling of orbits by frequency. In a sense the theorem says not to ask what happens to the orbit with a particular initial condition as a system is perturbed, but rather to consider the properties of an orbit with the same frequency.

Thus the KAM theorem says that most invariant circles (labeled by their frequency) persist for sufficiently small perturbations; however, in the proof of the theorem, "small" is indeed very small. In order to obtain better estimates for the domain of existence of invariant circles, it is better to ask about the existence of one particular circle instead of all smooth ones: the domain of existence of invariant circles in the space of smooth area-preserving maps is undoubtedly nothing like the simple ball assumed in the proof. For example, Herman (1985) has shown that there is at least one invariant circle (with $\omega = \gamma$) of the standard map when $k \le 0.029$. A computer-assisted version of this theorem, using interval arithmetic, attains the bound $k \le 0.91$ (de la Llave and Rana, 1990).

As we shall see in the next section, it is often easier to ask the converse question: when do rotational invariant circles not exist?

IV. INVARIANT CIRCLES

Though the KAM theorem gives us some insight into the existence and structure of invariant circles, its utility is limited because it is a perturbative result. There are, however, several important nonperturbative results about invariant circles for twist maps, i.e., maps of the cylinder that satisfy (1.30). In this section we prove Birkhoff's theorem, which implies that any rotational invariant circle must be the graph of a function, $y = Y(x)$. This theorem leads to techniques for proving the nonexistence of invariant circles—converse KAM theory; it implies that irregular components must be bounded by invariant circles; and it implies the existence of orbits that cross any region not containing invariant circles, thus showing that invariant circles are the only structures that prevent transport.

A. Rotational invariant circles

Let T be an area-preserving map on the cylinder. We suppose it is also end preserving: points with arbitrarily large positive y are mapped to similar points. This is the only possible case if the map arises from a Poincaré section of a flow, since the flow provides a smooth connection of the map to the identity.

An invariant circle is a curve $\mathcal{C}$ such that $T\mathcal{C} = \mathcal{C}$. A *rotational invariant circle* (RIC) is a closed loop that encircles the cylinder (i.e., is homotopically nontrivial; see Fig. 28). An invariant circle divides the cylinder into two invariant regions. To see this, consider the iterate of the region below an RIC. Since the map is continuous this iterate is a connected region. Since the circle is invariant, and the map is one to one, the iterate must have the circle as its boundary (otherwise it would have to "fold"). Finally, since the map is end preserving, points far below must retain below—thus the entire region must remain below. So a rotational invariant circle provides an absolute barrier to motion. Similarly, the region inside any invariant circle that encloses a finite area must remain inside.

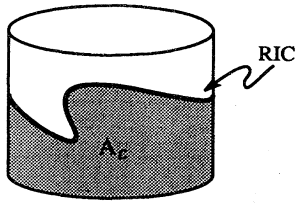

FIG. 28. Rotational invariant circle. An invariant loop that cannot be contracted to a point is a RIC.

B. Net flux

For a rotational circle $\mathcal{C}$, let $A_{\mathcal{C}}$ be the "algebraic area below" $\mathcal{C}$, that is, the value of the integral

$$A_{\mathcal{C}} = \int_{\mathcal{C}} y\, dx \ . \tag{4.1}$$

If y is positive on $\mathcal{C}$, then $A_{\mathcal{C}}$ is simply the geometric area between $\mathcal{C}$ and the circle $y=0$; if, however, the circle dips below $y=0$, then the contribution to Eq. (4.1) from this segment is negative, and the algebraic area differs from the geometric area. Moreover, in (4.1) we need not assume that the circle can be represented as a graph; so strictly speaking we should write the curve $\mathcal{C}$ in parametrized form as $(x(\lambda),y(\lambda))$ and integrate over λ.

The *net flux* is the area contained between a rotational circle $\mathcal{C}$ and its iterate $T\mathcal{C}$:

$$\mathcal{F}_N = A_{T\mathcal{C}} - A_{\mathcal{C}} \ . \tag{4.2}$$

When $T\mathcal{C}$ is above (below) $\mathcal{C}$ the contribution to (4.2) is positive (negative), as in Fig. 29. The net flux is independent of the choice of $\mathcal{C}$. To see this, choose a second curve $\mathcal{D}$. Because T is area preserving, the area contained between $\mathcal{C}$ and $\mathcal{D}$ is invariant; thus

$$A_{\mathcal{C}} - A_{\mathcal{D}} = A_{T\mathcal{C}} - A_{T\mathcal{D}} \ .$$

Rearranging this gives

$$A_{T\mathcal{D}} - A_{\mathcal{D}} = A_{T\mathcal{C}} - A_{\mathcal{C}} \ ;$$

so the net flux through $\mathcal{D}$ is the same as that through $\mathcal{C}$.

A map with zero net flux is *exactly* symplectic (recall Sec. I.C). We have already seen that the standard map has zero net flux in Eqs. (2.25) and (2.26).

A map that has an RIC must have zero net flux, since the net flux through the RIC is zero. This is why the condition (3.16) was required for the KAM theorem.

C. Birkhoff's theorem

Birkhoff showed that any invariant set U that looks like "half a cylinder" has a boundary that is the graph of some function $Y(x)$. In this section we sketch the proof of this theorem, and in the next section we discuss some of the practical consequences. Formally we have the

Theorem *(Birkhoff, 1920; Herman, 1983; Mather, 1984). Suppose T is a C^1 area-preserving, end-preserving twist map on the cylinder. Let U be an open invariant set homeomorphic to the cylinder such that there are $a<b$ satisfying*

$$\{x,y : y<a\} \subset U \subset \{x,y : y<b\} \ .$$

Then the boundary of U (∂U) is the graph $\{x,Y(x)\}$ of some continuous function Y.

The region U includes all points below $y=a$, is contained in the region $y<b$, and can have no holes. The point of the theorem is that ∂U cannot have any "whorls," for example, like those of a breaking wave. In particular, any continuous rotational invariant circle can be used as an upper boundary to form U; so the theorem implies that all RIC's are graphs.

1. Accessible points

The proof uses the concept of accessible points. Let $\gamma(t)=(x(t),y(t))$ be a curve embedded in U (γ cannot cross itself) and parametrized by t so that $y(-\infty)\to-\infty$. The deviation of γ from the vertical is defined to be the angle δ between a tangent to γ and the vertical. For those points $\gamma(t)$ such that $y(t)>y(t')$ for all $t'<t$, choose δ in the range $[-\pi/2,\pi/2]$; otherwise the branch of δ is chosen to make the deviation a continuous function (see Fig. 30).

A curve γ^R is tilted to the right if $\delta\le 0$ everywhere; i.e., its deviation from the vertical is everywhere to the right. Left-tilting curves are denoted γ^L.

As sketched in Fig. 31, a point $z_0\in U$ is *right accessible* if there exists a $\gamma^R\in U$ such that $\gamma^R(t_0)=z_0$.

2. Proof

A curve γ^R which tilts to the right is mapped onto another such curve by T. For example, suppose the angle δ at z is in the range $[-\pi,0]$; see Fig. 32. A vector v at z is mapped to Mv, where M is the linearization (1.31) of T at z. In particular, the twist condition (1.30) implies that the vertical at z is mapped to a right-tilting vector with tilt θ in the range $[-\pi,0]$. Since T preserves orientation, the angle δ' between Mv and the tangent to $T(\gamma^R)$ at $T(z)$ must be in the range $[-\pi,0]$. The deviation of

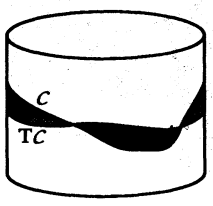

FIG. 29. Net flux through $\mathcal{C}$, shown by the difference between areas of the black region and the grey region.

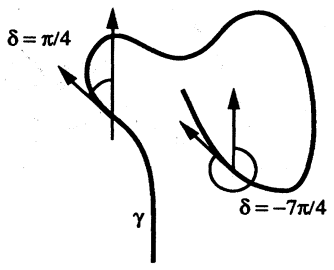

FIG. 30. Deviation from the vertical, δ.

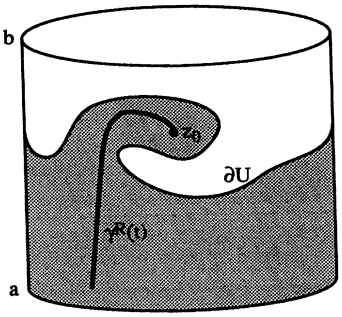

FIG. 31. Right accessible point.

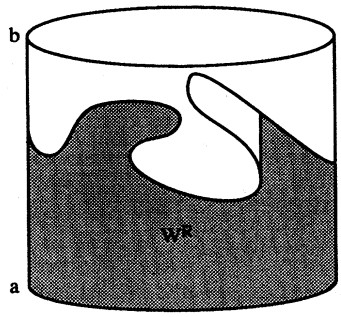

FIG. 33. Right accessible region, W^R.

$T(\gamma^R)$ from the vertical is the sum of these two angles and therefore must be to the right.

Let W^R and W^L be the subsets of U that are right and left accessible, respectively. The boundary of W^R consists of portions of ∂U together with vertical segments bounding those parts of U not right accessible (see Fig. 33). Since every point in W^R is on a curve that tilts to the right, W^R is mapped into itself by T:

$$T(W^R)\subset W^R .$$

Similarly, $T^{-1}(W^L)\subset W^L$ since T^{-1} twists to the left.

In fact, since T is area preserving and has zero net flux, $W^R=U$. If we suppose the contrary, then there is some portion of U that is not right accessible and is therefore a "lobe" bounded by a vertical on the right. Upon iteration any vertical tilts to the right, and therefore some portion of this lobe is mapped into W^R (Fig. 34). Now consider a circle $y=y_0$ far below ∂U. Since ∂U is contained between $y=a$ and $y=b$, the area of U above y_0 is finite. Furthermore, area preservation implies that the area of W^R above $y=y_0$ is mapped into a region with the same area. However, since the net flux through $y=y_0$ is zero, this gives a contradiction. Similarly, since T^{-1} twists to the left, $W^L=U$.

Thus every point of U is both right and left accessible, hence is vertically accessible. Therefore there exists a function $y=Y(x)$ describing ∂U.

D. Corollaries

1. Lipschitz corollary

A function $Y(x)$ is Lipschitz if there are finite slopes S_- and S_+ such that

$$S_+ \geq \frac{Y(x_1)-Y(x_0)}{x_1-x_0} \geq S_- \tag{4.3}$$

for all x_1 and x_0. These constants give a *Lipschitz cone*, which contains the graph of the function (see, e.g., Fig. 35). A Lipschitz function is continuous and differentiable almost everywhere.

A corollary of Birkhoff's theorem is that the function $Y(x)$ is Lipschitz. In fact, we can obtain explicit bounds on the slopes of an RIC. Upon iteration a vertical vector $\delta z=(0,\delta y)$ becomes $\delta z'=(\delta x',\delta y')=M\delta z$, which has the slope

$$S=\left.\frac{\delta y'}{\delta x'}\right|_x=\frac{\partial y'}{\partial y}\left[\frac{\partial x'}{\partial y}\right]^{-1} . \tag{4.4}$$

According to Eq. (1.30) the denominator of (4.4) is bounded below by the twist constant K; therefore there is a maximum slope, S_+. Inverse iteration of the vertical, using Eq. (1.32), leads to a minimum slope S_-:

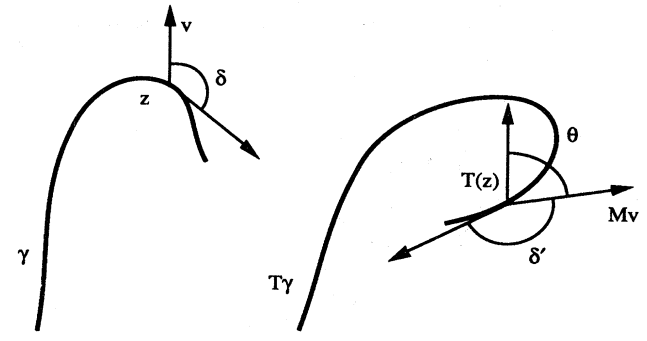

FIG. 32. Tilt property. The iterate of a curve that tilts to the right also tilts to the right.

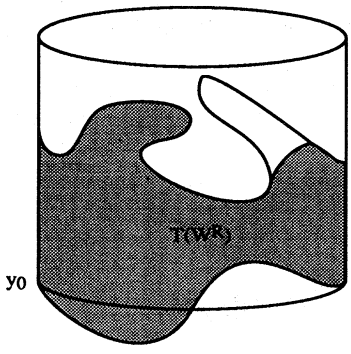

FIG. 34. Contradiction for the proof of Birkhoff's theorem. If $W^R\neq U$, then there are left-going lobes, and the iterate of W^R is strictly contained in W^R, violating area preservation.

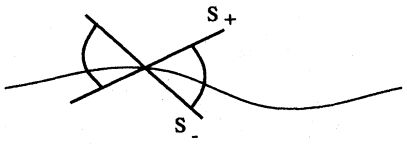

FIG. 35. Lipschitz cone with slopes S_+ and S_-.

$$S_- = \min\left[\frac{\partial y}{\partial y'}\left[\frac{\partial x}{\partial y'}\right]^{-1}\right]$$
$$= \min\left[-\frac{\partial x'}{\partial x}\left[\frac{\partial x'}{\partial y}\right]^{-1}\right]. \quad (4.5)$$

Since a rotational invariant circle intersects each vertical line exactly once, it must also intersect the iterate of each vertical exactly once. Thus the slopes S_+ and S_- bound the slope of the RIC.

2. Confinement corollary

Suppose the orbits of all points $y<a$ stay below some point b. Then there exists a rotational invariant circle between a and b.

To see this, we construct the set U for application of Birkhoff's theorem as follows. The iterates of all the points $y<a$ form an invariant set, which by assumption is contained below $y=b$; we shall call this set of iterates V. However, V cannot be used for Birkhoff's theorem because it is not necessarily homeomorphic to the cylinder (there will typically be lots of holes in the annulus $a<y<b$ corresponding to elliptic island chains). However, since V is below $y=b$, its complement has a connected component that contains all points $y>b$. Thus the complement of this connected component is a set contained below $y=b$, which satisfies the hypothesis of Birkhoff's theorem; we shall call this set U. The boundary of U is the RIC.

3. Converse KAM theory

Birkhoff's theorem leads to several criteria for the nonexistence of invariant circles which have varying effectiveness in practice.

(a) *Climbing orbits.* If there is an orbit that climbs arbitrarily far up the cylinder, then there are no rotational invariant circles. More precisely, consider an annulus $a<y<b$. If there is an orbit going from below this annulus to above it, then there are no RIC's contained in the annulus. Furthermore, since RIC's must be Lipschitz, for any point z there is an annulus, with height

$$S_+ + |S_-| ,$$

inside of which any RIC containing z must lie. In practice this criterion is not too useful, since even when RIC's do not exist it may take many iterations for orbits to climb even a small distance.

(b) *Heteroclinic connections.* Suppose the unstable manifold of some periodic orbit intersects the stable manifold of another. Then there can be no RIC's contained between them. This could be a practical criterion because the stable and unstable manifolds can be computed numerically. Furthermore, this is really what underlies the resonance overlap criterion, Sec. II.C.

(c) *Lipschitz criteria.* Using the Lipschitz bounds on slopes, one can obtain restrictive criteria for the nonexistence of RIC's. Consider the iteration of a small vertical vector $\delta z_0=(0,1)$ at the point $z_0=(x_0,y_0)$, using the linear map (1.30). Upon one iteration we obtain

$$\delta z_1=(\delta x_1,\delta y_1)=M_{z_0}\begin{bmatrix}0\\1\end{bmatrix}=\left[\frac{\partial x'}{\partial y},\frac{\partial y'}{\partial y}\right],$$

which has positive δx_1 by (1.30). However, a second iteration gives

$$\delta z_2=M_{z_1}\delta z_1 \Longrightarrow \delta x_2=\frac{\partial x''}{\partial x'}\frac{\partial x'}{\partial y}+\frac{\partial x''}{\partial y'}\frac{\partial y'}{\partial x}. \quad (4.6)$$

If $\delta x_2<0$ there can be no RIC through z_0, because the orbit of z_0 would have to be on the circle and it could not be a graph (see Fig. 36). For example, for the standard map, (1.36), (4.6) becomes

$$\delta x_2=2-k\cos(2\pi x') . \quad (4.7)$$

Now, since a RIC must intersect every vertical, if (4.7) is negative for any x', there are no rotational invariant circles. Thus when $|k|>2$, there are no RIC's for the standard map. Mather (1984) refines this criterion using the explicit Lipschitz cone to obtain the bound $|k|>4/3$. MacKay and Percival (1985) use a further refinement of this criterion to obtain the bound $|k|>63/64$. They utilize the computer to obtain this result: each floating-point calculation is given explicit bounds so that the result is rigorous. Furthermore, Stark (1986) has shown that the criterion of MacKay and Percival is exhaustive: if there is no invariant circle, the method will eventually show nonexistence. These bounds compare favorably with the result (2.22) of Greene.

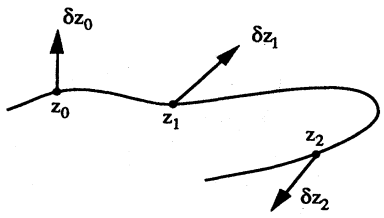

FIG. 36. Converse KAM criterion. If the iterate of a vertical vector eventually turns around, there is no RIC.

V. VARIATIONAL PRINCIPLES

In this section we show that any twist map has a Lagrangian variational principle. This variational formulation turns out to be of great utility. In Secs. VI and VII we shall discuss the theory of Aubry and Mather, which uses this formulation to classify what one could call the "regular" orbits of a twist map. The variational principle also has a physical interpretation in terms of phase-space areas, and in Sec. VIII we show that it provides compact and computationally efficient formulas for the area of resonances and escaping fluxes.

A. Generating function

Let $T:(x,y)\to(x',y')$ be the lift (as discussed in Sec. II.A) of a twist mapping to the plane. We shall show there exists a generating function $F(x,x')$ such that

$$\begin{aligned} y &= -F_1(x,x') , \\ y' &= F_2(x,x') , \end{aligned} \tag{5.1}$$

or, alternatively, a "one-form" dF:

$$dF(x,x') = y'\,dx' - y\,dx . \tag{5.2}$$

Here the subscripts indicate derivatives with respect to the first and second arguments, respectively. F is a generating function for a canonical transformation (F_1 in Goldstein's notation).

To show the existence of F we must first invert the relation $x'(x,y)$ to obtain $y(x,x')$; we do this geometrically. Consider the verticals $x=\xi$ and $x=\xi'$ in the plane. The curve $T(x=\xi)$ intersects ξ' exactly once by the twist condition. Define $y'(\xi,\xi')$ to be this intersection (Fig. 37). Similarly, define $y(\xi,\xi')$ to be the unique intersection of $T^{-1}(x=\xi')$ with the vertical $x=\xi$.

Using these functions, we define the generating function by integration,

$$F(x,x') = \int_\gamma^{(x,x')} y'(\xi,\xi')\,d\xi' - y(\xi,\xi')\,d\xi , \tag{5.3}$$

where γ is a path (see, e.g., Fig. 38), which begins at some arbitrary point and ends at (x,x'). In fact, Eq. (5.3) is independent of the choice of path. To see this consider a second path $\bar\gamma$ that has the same end points as γ. By Stokes's theorem the integral $\oint y(x,x')\,dx$ around the closed loop $\gamma-\bar\gamma$ is the integral of the area enclosed: $\int dy\wedge dx$. Since (x',y') is the iterate of (x,y), area preservation implies that this is the same as $\int dy'\wedge dx' = \oint y'(x,x')\,dx'$ over this same loop. Thus the integrals of dF along γ and $\bar\gamma$ are equal (we say that dF is an *exact* one-form in the plane).

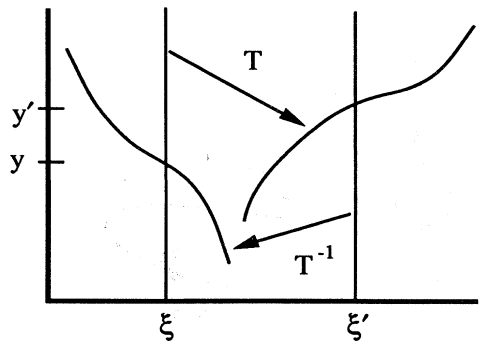

FIG. 37. Definition of $y(x,x')$ and $y'(x,x')$.

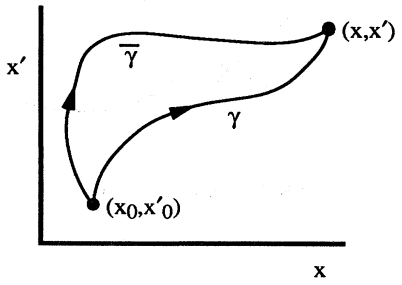

FIG. 38. Curves γ and $\bar\gamma$ in (x,x') space.

By construction, the derivative of F with respect to its first argument is $-y(x,x')$ and with respect to its second is $y'(x,x')$, as required.

The twist condition (1.30) translates into a requirement upon the second derivative of F. Using (1.37) we obtain

$$F_{12}(x,x') = -\frac{\partial y}{\partial x'} = -\left[\frac{\partial x'}{\partial y}\right]^{-1} \le -\frac{1}{K} < 0 ; \tag{5.4}$$

so the mixed second partial derivative of F is negative-definite.

The mapping generated by F is area preserving because $dy' = F_{12}dx + F_{22}dx'$ and $dy = -F_{11}dx - F_{12}dx'$ imply that the two area elements

$$\begin{aligned} dy\wedge dx &= -F_{12}dx'\wedge dx , \\ dy'\wedge dx' &= F_{12}dx\wedge dx' = -F_{12}dx'\wedge dx \end{aligned} \tag{5.5}$$

are the same.

Finally, this construction provides a useful interpretation of the generating function. Consider a curve $\mathcal{C}$ and its iterate $\mathcal{C}'$ (see Fig. 39). The area under $\mathcal{C}$ is the integral $\int y\,dx$ along $\mathcal{C}$, while that under $\mathcal{C}'$ is $\int y'\,dx'$.

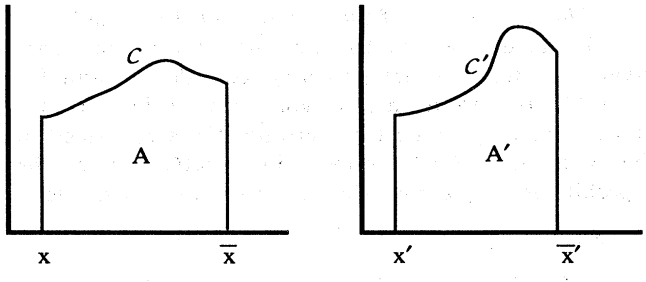

FIG. 39. Area under a curve and its iterate.

The difference between these areas is

$$A'-A=\int_{\mathcal{C}'}y'\,dx'-\int_{\mathcal{C}}y\,dx$$
$$=F(\bar{x},\bar{x}')-F(x,x')\ , \tag{5.6}$$

where we recall Eq. (5.2). We shall discuss and rederive this relationship in Sec. VIII.

B. Net flux

The net flux across a rotational circle $\mathcal{C}$ is the difference between the area under $\mathcal{C}$ and that under $T\mathcal{C}$ (recall Sec. IV.B). A rotational circle is a curve that ranges from (x,y) to $(x+1,y)$. Since the mapping is periodic, $T\mathcal{C}$ ranges from (x',y') to $(x'+1,y')$. Using these curves in Eq. (5.6) gives a formula for the net flux:

$$\mathcal{F}_N=F(x+1,x'+1)-F(x,x')\ . \tag{5.7}$$

The net flux is zero if the generating function is a periodic function of $\frac{1}{2}(x+x')$; it can depend arbitrarily on $x'-x$. Such a mapping is called *exactly symplectic* because in this case the one-form $y'\,dx'-y\,dx$ is exact on the cylinder: its integral is path independent even for paths that encircle the cylinder.

C. Examples

1. Standard map

A generating function for the standard map (1.36) is

$$F(x,x')=\frac{1}{2}(x-x')^2-V(x)\ ,$$
$$V(x)=-\frac{k}{4\pi^2}\cos(2\pi x)\ . \tag{5.8}$$

This is the same as the energy per site for the Frenkel-Kontorova model (1.39).

From another point of view the generating function is a discrete version of the Lagrangian for a dynamical system. For the standard map, it has the familiar form of kinetic minus potential energies, where the "velocity" is $x'-x$ for the discrete-time system, and the potential is $V(x)$. Thus we see that the standard map is a discrete approximation to the pendulum.

Equation (5.8) confirms that the standard map has zero net flux, by Eq. (5.7).

2. Billiards

The generating function for a convex billiard is the function that gives the length between two boundary points. Let (X,Y) represent rectangular coordinates in the plane of the billiard. Using Birkhoff coordinates $(s,\cos\theta)$, Sec. I.F, we see that the generating function is

$$F(s,s')=\{[X(s)-X(s')]^2+[Y(s)-Y(s')]^2\}^{1/2}\ , \tag{5.9}$$

where $(X(s),Y(s))$ represents the billiard boundary. The derivatives of F are

$$\frac{\partial}{\partial s}F(s,s')=\frac{1}{F}\left[\frac{\partial X}{\partial s}[X(s)-X(s')]+\frac{\partial Y}{\partial s}[Y(s)-Y(s')]\right]=-\cos\theta\ ,$$
$$\frac{\partial}{\partial s'}F(s,s')=\cos\theta'\ , \tag{5.10}$$

since the vector $(\partial X/\partial s,\partial Y/\partial s)$ is the unit tangent to the boundary (recall Fig. 10). This confirms again that the momentum coordinate is $\cos\theta$. In these coordinates the billiard map is area preserving.

The twist for the billiard is

$$F_{12}(s,s')=\frac{\sin\theta\sin\theta'}{F}\ . \tag{5.11}$$

Since, for a convex billiard, $0<\theta<\pi$, the mapping has twist; however, it twists to the left, since $F_{12}\geq 0$. Therefore the sign convention for billiards is opposite to that which is used in this paper. To translate our discussion in Secs. VI–VIII for billiards, replace "minimizing" by "maximizing."

The circle billiard has the generating function

$$F=2r\sin\left[\frac{s'-s}{2r}\right]\ , \tag{5.12}$$

where r is the radius. Since F is a function only of $s'-s$, the circle billiard is trivially integrable: the momentum is conserved. Obtaining a generating function for more general billiards, such as an ellipse (which is also integrable) or the stadium (Bunimovich, 1974), is left as an exercise!

D. Action

For a continuous-time Lagrangian system, the action is the integral of the Lagrangian $L(q,\dot{q},t)=p\dot{q}-H(p,q,t)$ along a path $q(t)$ in configuration space [recall Eq. (1.5)]. Orbit segments $q(t)$ of this dynamical system are stationary points of the action with respect to variations with given end points $q(t_0)=x$, and $q(t_1)=x'$. We shall show here that the value of this action on the orbit is the function $F(x,x')$, which generates the mapping for this flow. Suppose the Lagrangian depends periodically on time, and without loss of generality suppose the period is 1. Let $q(t)$ be an orbit segment for one period, and define

$$F(x,x')=\int_{\substack{0\\ q(t)\text{ stationary}}}^{1}L(q,\dot{q},t)\,dt\ . \tag{5.13}$$

Now (5.13) is the value of the action of the exact orbit from $q(0)=x$ to $q(1)=x'$, and the mapping generated by F is the time 1 mapping of the Lagrangian flow. To obtain the action for several periods, we merely have to sum (5.13) over the intermediate steps:

$$W\{x_m,x_{m+1},\ldots,x_n\}=\sum_{t=m}^{n-1}F(x_t,x_{t+1})\ . \tag{5.14}$$

This is the same as the action (1.5), restricted to a path that is an exact orbit for each period; it depends only on the configuration points at the discrete times $t=n$.

An *orbit segment* is a configuration $\{x_m,\ldots,x_n\}$ that is a stationary point of the action holding x_m and x_n fixed. Setting the variation of the action to zero gives the equations

$$\delta W=0\Longrightarrow\frac{dW}{dx_j}=F_2(x_{j-1},x_j)+F_1(x_j,x_{j+1})=0 \tag{5.15}$$

for $m<t<n$, which implies that the two definitions of momentum (5.1) agree at each point on the orbit:

$$y'(x_{j-1},x_j)=y(x_j,x_{j+1})=y_j\ . \tag{5.16}$$

An (m,n) periodic orbit, (2.10), is determined by the action

$$W_{(m,n)}\{x_0,x_2,\ldots,x_{n-1}\}=\sum_{t=0}^{n-1}F(x_t,x_{t+1})|_{x_n=x_0+m}\ , \tag{5.17}$$

which is a function of the $n-1$ distinct points on the orbit. The (m,n) periodic orbit is a stationary point of $W_{(m,n)}$ upon variation of all its arguments. This yields the same equations as before when $0<t<n$. Variation with respect to x_0 gives the equation

$$F_1(x_0,x_1)+F_2(x_{n-1},x_n)=0\ ,$$

which implies the periodicity condition $y_n=y_0$.

An *orbit* is a bi-infinite sequence $\{\ldots,x_{t-1},x_t,x_{t+1},\ldots\}$ such that every finite subsequence is an orbit segment. Thus the action $W\{x\}$ is stationary for each t.

For example, for the standard map, stationary points of the action must satisfy Eq. (1.38), which is the Lagrangian form of the equations (1.36). Similarly, for the billiard, Eq. (5.16) implies that the angle of incidence equals the angle of reflection for each bounce.

VI. PERIODIC ORBITS

In this section and the next we shall discuss the theory of Aubry and Mather, which shows the existence of minimizing and minimax orbits for each frequency ω for an area-preserving twist mapping. In the process, many properties of these orbits will become clear.

A. Minimizing orbits

The action of an orbit was defined in Eq. (5.14). Its first variation on an orbit is zero according to Eq. (5.15). This implies that the action does not change under an infinitesimal variation of the configuration to first order: $\delta W\{x\}=0$. The second variation of the action about an orbit is not generally zero. Consider first a finite segment of an orbit, $\{x_m,\ldots,x_n\}$; let $\delta^2W\{x_m,\ldots,x_n\}$ be the quadratic form obtained from the second-order term in the expansion of W for fixed x_m and x_n:

$$\delta^2W\{\delta x\}=\sum_{j,k=m+1}^{n-1}\delta x_j\frac{\partial^2W}{\partial x_j\partial x_k}\delta x_k\ . \tag{6.1}$$

An orbit segment is *locally minimizing* if δ^2W is nonnegative for all vectors $\{\delta x_{m+1},\ldots,\delta x_{n-1}\}$. If δ^2W is positive-definite, then the minimum is nondegenerate.

The orbit corresponding to the infinite sequence $\{\ldots,x_m,\ldots x_n,\ldots\}$ is defined to be locally minimizing if *every finite* segment is locally minimizing.

Consider now arbitrary variations $\{\xi_m,\ldots,\xi_n\}=\{x_m,x_{m+1}+\delta x_{m+1},\ldots,x_{n-1}+\delta x_{n-1},x_n\}$ about some orbit segment $\{x\}$ with fixed end points (here the δx_i's can be of arbitrary size). An orbit segment is defined to be *minimizing* if for every variation $\{\xi\}$

$$W\{\xi\}-W\{x\}\geq 0\ . \tag{6.2}$$

If *every finite* segment of an orbit is minimizing, then the orbit is minimizing. In this definition it is important to allow only variations with compact support; otherwise the action difference $W\{\xi\}-W\{x\}$ would not necessarily be finite (being an infinite sum), and the two orbits could not be compared. Furthermore, anchoring the asymptotic ($t\to\pm\infty$) behavior of the orbit acts as a kind of boundary condition, and we shall find different minimizing orbits when different boundary conditions are imposed.

It is not obvious that minimizing orbits exist. We shall first show that there are (m,n) minimizing periodic orbits for any frequency. There are two steps in this demonstration: first we consider orbits that minimize $W_{(m,n)}$, and then we show that these also minimize W. In Sec. VII we shall consider irrational ω.

B. Existence of (m,n) orbits

The Poincaré-Birkhoff theorem implies that a twist mapping has at least two periodic orbits for each (m,n). Actually, this theorem applies to a more general class of maps: maps on an annulus that preserve the two boundaries, rotating them in opposite directions. Such maps need not satisfy the twist condition (the two ends of a vertical line must move in opposite directions, but the intermediate points are unconstrained). To prove his theorem, Birkhoff used intricate geometric arguments (Birkhoff, 1913). For the twist case the existence of these orbits follows more simply from the variational principle. The first orbit appears as a minimum of $W_{(m,n)}$, and the second will follow from the minimax principle. The proof of the existence of a minimum is based on the

Growth condition. *For an area-preserving twist mapping with zero net flux the generating function is bounded by*

$$F(x,x') \geq A - B|x-x'| + C|x-x'|^2 , \tag{6.3}$$

where B and C are positive.

Proof. Let $\xi_\lambda = x + \lambda(x'-x)$ represent the line connecting x to x' as λ ranges from 0 to 1. Then for any function $F(x,x')$ we have the identity

$$F(x,x') = F(x,x) + \int_0^1 d\lambda \, F_2(x,\xi_\lambda)(x'-x) . \tag{6.4}$$

Repeating this construction on F_2 gives

$$F(x,x') = F(x,x) + \int_0^1 d\lambda \, F_2(\xi_\lambda,\xi_\lambda)(x'-x) - \int_0^1 d\lambda \int_0^\lambda d\mu \, F_{12}(\xi_\mu,\xi_\lambda)(x'-x)^2 . \tag{6.5}$$

Define $A = \min[F(x,x)]$ and $B = \max|F_2(x,x)| > 0$. These exist by periodicity (5.7) when the net flux is zero. Finally, let $C = \frac{1}{2}K > 0$, where K is the twist constant (5.4). Substituting these into the integrands in (6.5) gives (6.3) directly.■

This result can be generalized to $2N$-dimensional symplectic twist maps. Here we assume that the mapping has uniformly positive-definite twist: there exists a positive-definite matrix $\mathbf{C}$, such that for any vector $\mathbf{v}$, $\mathbf{v}\cdot F_{12}(\mathbf{x},\mathbf{x}')\cdot\mathbf{v} \leq -\mathbf{v}\cdot\mathbf{C}\cdot\mathbf{v}$ for all $\mathbf{x},\mathbf{x}'$. For this case the above proof of the growth condition can be directly transcribed (MacKay *et al.*, 1989).

Using the growth condition in the action for a periodic orbit, we can prove the

Poincaré-Birkhoff theorem. *For an area-preserving twist mapping with zero net flux there is a periodic orbit for each (m,n).*

Proof. We shall obtain the orbit as a global minimum of $W_{(m,n)}$ [see Eq. (5.17)]. $W_{(m,n)}$ is a function on the space of periodic configurations $\{x_0,x_1,\ldots,x_{n-1}\} \in \mathbb{R}^n$. Since the mapping is periodic, we can, without loss of generality, choose x_0 to lie in the interval [0,1]; so the space of configurations reduces to $[0,1]\times\mathbb{R}^{n-1}$. To guarantee that $W_{(m,n)}$ has a minimum, we must find a compact subset on which $W_{(m,n)}$ is bounded. By (6.3) $W_{(m,n)}$ satisfies the bound

$$W_{(m,n)} \geq nA + \sum_{j=0}^{n-1} (-B|x_{j+1}-x_j| + C|x_{j+1}-x_j|^2) . \tag{6.6}$$

In particular, $W_{(m,n)} \geq n(A - \frac{1}{4}B^2/C)$ is bounded from below.

Now consider the set of configurations for which $W_{(m,n)} \leq nA + D$, for some constant D. We can show that this is a compact set in the space of configurations. In particular, the bound on $W_{(m,n)}$ implies that the sum in (6.6) is smaller than D; therefore each term is bounded. This implies that $|x_{j+1}-x_j|$ is bounded, and therefore, since $x_0 \in [0,1]$, $|x_t - x_0|$ is bounded. Thus each of the x_t for $0 < t < n$ is bounded.

Outside the compact set $W_{(m,n)}$ is large. On the other hand, since $W_{(m,n)}$ is bounded below on the compact set it must have a minimum. Thus there exists an (m,n) periodic orbit that minimizes $W_{(m,n)}$.■

The minimum is not unique. For example, if $\{x_0,x_1,\ldots,x_{n-1}\}$ is a minimum, then so is any translate:

$$\{x_j+k, x_{j+1}+k, \ldots, x_{n-1}+k, x_0+k, \ldots, x_{j-1}+k\} \tag{6.7}$$

for any j where the integer k is chosen so that $x_j + k$ is in the unit interval [0,1]. Thus there are at least n minima in the domain $[0,1]\times\mathcal{R}^{n-1}$.

As an example, consider the (0,1) orbit for the standard map (5.8). The action is

$$W_{(0,1)}\{x_0\} = -V(x_0) , \tag{6.8}$$

which is a periodic function and therefore has at least one minimum. For the standard map this occurs at $x = \frac{1}{2}$, corresponding to the hyperbolic fixed point. Note that minima of W correspond to maxima of V and therefore to dynamically unstable orbits.

In exceptional cases, such as the integrable twist map (2.1), there is an entire curve of minima, forming the rational frequency invariant circle. In this case Eq. (5.8) yields the action

$$W_{(m,n)} = \frac{1}{2}\sum_{t=0}^{n-1} |x_t - x_{t+1}|^2 . \tag{6.9}$$

The extrema that satisfy the constraint $x_n = x_0 + m$ are the orbits $x_t = x_0 + mt/n$ for any choice of x_0. For each x_0, $W_{(m,n)}$ has the same value; since these are the only extrema, a variation about these orbits can never decrease the action. Thus these extrema are degenerate minima.

C. Aubry's fundamental lemma

We have obtained periodic orbits that minimize the periodic action. To show that these orbits are minimizing, we need the "fundamental lemma" of Aubry (1983b). The point is that even though we have shown that the periodic configuration minimizes $W_{(m,n)}$, it is possible that a variation which is not (m,n) periodic will decrease the action of the infinite orbit, W.

The fundamental lemma utilizes the twist condition and its concomitant distinction between the x and y coordinates as an essential hypothesis. First, we use the fact that an orbit is determined entirely by its configuration sequence, as in Eq. (5.14). Furthermore, we shall see that there is a frequency (or mean velocity) associated with each minimizing orbit, and that if one minimizing orbit is moving more rapidly through phase space than another,

then the paths of the orbits must diverge—one cannot oscillate about the other. The twist condition necessarily orders these orbits in y: larger momentum means larger frequency.

Aubry's lemma is related to Morse's theorem (Morse, 1924) for geodesics (two minimum length geodesics on a toroidal surface cross at most once) and is a global version of a theorem in the calculus of variations (locally minimizing orbits have no conjugate points; see Gelfand and Fomin, 1963). We prove only the simplest version of this lemma (MacKay and Stark, 1985):

Aubry's fundamental lemma. *Let $\{x\}$ and $\{\xi\}$ be two distinct minimizing orbits. Then they cross at most once.*

To define the crossing of orbits, draw the orbits in the (t,x) plane and join successive points with straight lines to form the continuous curves

$$x(t)=(x_j-x_{j-1})(t-j)+x_j \quad \text{for } j-1\le t\le j \ . \tag{6.10}$$

Similarly, construct the curve for ξ. The orbits $\{x\}$ and $\{\xi\}$ are said to *cross* if the function $x(t)-\xi(t)$ has a zero.

Proof of Aubry's fundamental lemma. Suppose the converse, that $\{x\}$ and $\{\xi\}$ cross twice. We shall obtain a contradiction. There are three possible cases: (i) The crossing points both occur at noninteger values of t, as in Fig. 40; (ii) one of them occurs at integer t; or (iii) they both occur at integer values of t.

Case (i). We construct deformations of $\{x\}$ and $\{\xi\}$ and show that at least one of these has smaller action, implying that not both $\{x\}$ and $\{\xi\}$ can be minimizing. Let the two trajectories cross between times j and $j+1$ and times k and $k+1$. Define the deformations

$$\begin{aligned}\{\tilde{\xi}\}&=\{\ldots,\xi_{j-1},\xi_j,x_{j+1},\ldots,x_k,\xi_{k+1},\xi_{k+2},\ldots\}\\ \{\tilde{x}\}&=\{\ldots,x_{j-1},x_j,\xi_{j+1},\ldots,\xi_k,x_{k+1},x_{k+2},\ldots\}\end{aligned} \tag{6.11}$$

as sketched in Fig. 41. Note that it is necessary to have $\{x\}$ and $\{\xi\}$ cross twice to construct these deformations, because the definition of minimizing required that the variation occur only on finite segments.

Consider the orbit segments running from time j to $k+1$. Since $\{x\}$ and $\{\xi\}$ were assumed to be minimizing, the new segments must not have smaller action. Adding the actions of these two segments we obtain

$$W\{\tilde{x}\}+W\{\tilde{\xi}\}=F(x_j,\xi_{j+1})+W\{\xi_{j+1},\ldots,\xi_k\}+F(\xi_k,x_{k+1})+F(\xi_j,x_{j+1})+W\{x_{j+1},\ldots,x_k\}+F(x_k,\xi_{k+1}) \ . \tag{6.12}$$

Subtracting from this the sum $W\{x\}+W\{\xi\}$ and regrouping terms gives

$$\begin{aligned}W\{\tilde{x}\}+W\{\tilde{\xi}\}-W\{x\}-W\{\xi\}&=F(x_j,\xi_{j+1})+F(\xi_j,x_{j+1})-F(x_j,x_{j+1})-F(\xi_j,\xi_{j+1})\\ &\quad+F(x_k,\xi_{k+1})+F(\xi_k,x_{k+1})-F(x_k,x_{k+1})-F(\xi_k,\xi_{k+1}) \ .\end{aligned} \tag{6.13}$$

Each of these sets of four terms can be shown to be negative. In general, consider two points (x,x') and (ξ,ξ') and assume that there is a crossing so that $(x-\xi)(\xi'-x')$ is positive; then

$$F(x,\xi')+F(\xi,x')-F(x,x')-F(\xi,\xi')=\int_{\xi}^{x}d\lambda\int_{x'}^{\xi'}d\mu\,F_{12}(\lambda,\mu)\le-\frac{1}{K}(x-\xi)(\xi'-x') \ ,$$

where K is positive by the twist condition (5.4). (In the case of no crossing, the inequality must be reversed because $d\lambda\,d\mu$ is negative.) Therefore after a crossing

$$F(x,\xi')+F(\xi,x')-F(x,x')-F(\xi,\xi')<0 \ ; \tag{6.15}$$

so the difference between the actions of the modified orbits and the original orbits satisfies

$$W\{\tilde{x}\}+W\{\tilde{\xi}\}-W\{x\}-W\{\xi\}<0 \ . \tag{6.16}$$

This contradicts the assumption that both $\{x\}$ and $\{\xi\}$ are minimal.

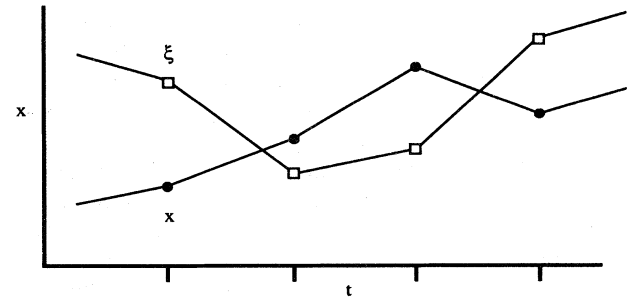

FIG. 40. Crossing configurations. The configurations $\{x\}$ and $\{\xi\}$ cross twice.

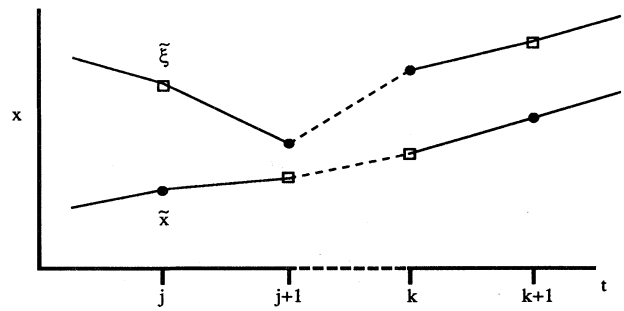

FIG. 41. Deformed configurations that no longer cross.

Case (ii) is proved similarly. The difference between the actions has contributions only from the noninteger crossing, but it still is negative.

Case (iii). Both crossing points are at integer times, say, $t=j+1$ and k. Choose the new segments as in Eq. (6.11). Now the sum of the actions of the new segments from j to $k+1$ is the same as for the old segments. However, the new segments cannot be stationary points of the action because, although $\tilde{\xi}_j=\xi_j$ and $\tilde{\xi}_{j+1}=\xi_{j+1}$, $\tilde{\xi}_{j+2}\neq\xi_{j+2}$; and stationarity uniquely determines ξ_{j+2}. Since the new segments are not even stationary they cannot be minimizing. This contradicts the assumption that the original orbits are minimizing, since the action is unchanged in value.■

We shall apply Aubry's fundamental lemma to determine various properties of minimizing configurations in the new few sections. One direct result is the

Corollary. *Two (m,n) minimizing orbits cannot cross.*

Proof. Suppose $\{x\}$ and $\{\xi\}$ are both minimizing (m,n) periodic orbits. Then they cannot cross, for if they cross once, then periodicity implies that they cross each period, and therefore infinitely often.■

Furthermore, minimizing (m,n) orbits have a monotonicity property. An orbit is said to be *monotone* if for any integers t,t',j and j'

$$x_t+j<x_{t'}+j' \Longrightarrow x_{t+1}+j<x_{t'+1}+j' \ . \tag{6.17}$$

Corollary. *Minimizing (m,n) orbits are monotone.*

Proof. Let $x_t+j\to x_t$ and $x'_t+j'\to\xi_t$ and apply the fundamental lemma: ξ_t cannot cross x_t.■

We shall see in Sec. VII that monotone orbits are ordered in the same way as simple rotations on the circle.

D. Minimizing (m,n) orbits

We now are set up to prove the existence of minimizing periodic orbits. We follow the discussion of Banget (1988) to prove the

Theorem *(Aubry and Le Daeron, 1983). For an area-preserving twist mapping there is a minimizing periodic orbit for every (m,n), where m and n are coprime.*

Proof. Let $\{x\}$ be the periodic extension of the configuration that minimizes $W_{(m,n)}$. We must show that there is no infinite configuration that has smaller action. For example, we consider an orbit $\{\xi\}$ of type (km,kn), which minimizes $W_{(km,kn)}$. By the fundamental lemma, this orbit cannot cross any of its translates. Now suppose $\{\xi\}$ is not also of type (m,n). Then $\xi_{t+n}\neq\xi_t+m$. Since $\xi_{t+n}-m$ does not cross ξ_t we must have either

$$\xi_{t+n}-m>\xi_t \quad \text{or} \quad \xi_{t+n}-m<\xi_t \ . \tag{6.18}$$

Consider the first case. Shifting time by n steps implies that $\xi_{t+2n}-m>\xi_{t+n}$, and therefore $\xi_{t+2n}-2m>\xi_{t+n}-m>\xi_t$. Repeating this k times gives $\xi_{t+kn}-km>\xi_t$. This is a contradiction, since we assumed it was of type (km,kn). So if an (m,n) minimizing periodic orbit has a smallest period n, then m and n are coprime.

We have just shown that if $\{x_0,x_1,\ldots,x_{n-1}\}$ minimizes $W_{(m,n)}$, then

$$\{x_0,x_1,\ldots,x_{n-1},x_0+m,x_1+m\ , \ldots,x_{n-1}+(k-1)m\} \tag{6.19}$$

minimizes $W_{(km,kn)}$ for all $k>1$. Since the segment (6.19) is minimal, its action must be less than that of any variation with the same end points. Since k is arbitrary, this implies that any variation of the orbit $\{x\}$ with compact support must increase the action of $\{x\}$. Thus $\{x\}$ is a minimal orbit.■

This theorem apparently cannot be generalized to higher dimensions. For the case of geodesic flow on a torus, where an analogous theorem holds in two dimensions, Hedlund (1932) has given a counterexample on a three-torus. The difficulty in this case is that there is no natural generalization of the idea of crossing: curves do not separate regions in a space with more than two dimensions. Thus it seems that minimizing orbits may not be as important in higher-dimensional systems, though orbits that minimize the periodic action may still play an important role (Kook and Meiss, 1989; Golé, 1990).

E. Minimax principle

The existence of a minimizing (m,n) orbit immediately implies the existence of another orbit, the minimax orbit. This occurs because the translates $\xi_t=x_{t+k}+j$ of a minimizing orbit are also minimizing; thus the existence of one minimum for $W_{(m,n)}$ implies directly that there are many minima. Between these minima there must be other critical points. The *Morse index* of a critical point of a function [i.e., a point for which $Df(x)=0$] is defined to be the number of downward directions of the function at that point (Milnor, 1963). Thus a minimizing orbit has index zero. The minimax principle, originally due to Birkhoff, implies that there is an orbit of index 1. To show this more formally we construct a new orbit by constrained minimization.

Theorem. *For every (m,n) there exists an index-1, periodic orbit—the minimax orbit.*

Proof. Any translate $\{\xi_t\}$ of $\{x_t\}$ does not cross $\{x_t\}$, since both orbits are minimizing. Choose the translate for which ξ_0 is closest to x_0. Now choose a path $\zeta(\lambda)=\{x_0(\lambda),\ldots,x_{n-1}(\lambda)\}$ for $\lambda\in[0,1]$ connecting

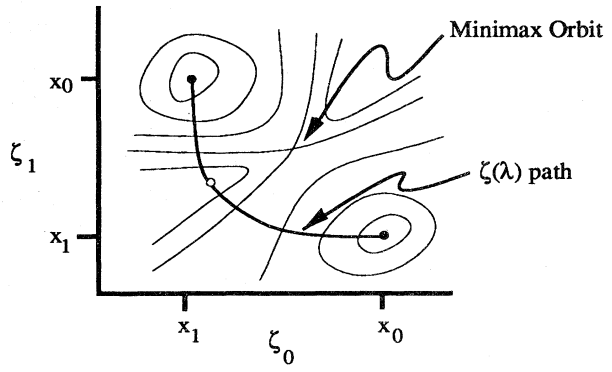

FIG. 42. Minimax construction for $n=2$. The minimax orbit occurs at a saddle point of W between neighboring minima.

these two neighboring minima of $W_{(m,n)}$. Since $W_{(m,n)}$ is continuous it must have a maximum along this path. In Fig. 42 we sketch the $n=2$ case. Minima occur at the points $\{x_0,x_1\}$ and $\{x_1,x_0\}$. The maximum along the path ζ is shown in the figure as the open circle. Now vary the path $\zeta(\lambda)$ to find the smallest of these maxima. This gives a critical point of $W_{(m,n)}$ and therefore an (m,n) periodic orbit. This minimum over the maxima is the minimax orbit.■

In addition, the minimax orbit is well ordered with respect to the minimizing orbit, in the sense of (6.17) (Mather, 1986).

The minimizing and minimax orbits form the "island chain" structure seen in Sec. II. In fact, one can see that the residue (2.16) of a nondegenerate minimizing orbit must be negative (MacKay and Meiss, 1983), indicating that it is hyperbolic (the orbit is parabolic if the minimum is degenerate). On the other hand, the residue of a nondegenerate minimax orbit must be positive, so that it is either elliptic or hyperbolic with reflection.

This is most easily seen for the fixed-point case. Here the action is $W_{(0,1)}=F(x,x)=-V(x)$. This is a periodic function, and so it necessarily has at least one minimum and one maximum. However, the minimum of W corresponds to the maximum of the potential energy and therefore gives the unstable orbit [as we saw in Eq. (2.17)]! Similarly, the minimax orbit sits at the minimum of V and is therefore elliptic. Remarkably, this circumstance generalizes to any (m,n) orbits.

When the minimax orbit is elliptic we have the familiar island-chain structure. If it is reflection hyperbolic, then this typically means that the elliptic orbit has undergone a period-doubling bifurcation (Greene *et al.*, 1981), signaling the destruction of most of the invariant circles in the island chain. Even in this case the unstable manifolds of the hyperbolic, minimizing orbit can be used to form the "separatrix" of an island, as we shall see in Sec. VIII.

Aside from the minimizing and minimax orbits, there are of course many other periodic orbits in a typical mapping. Some of these can be understood by techniques similar to the above. For example, the librating orbits within an island chain can be thought of as ordered orbits of the mapping T^n with respect to rotation about the minimax fixed point in the center of the island. Since T^n typically has twist in the neighborhood of such a point (Sec. I.F), the above theorems prove the existence of librating periodic orbits for all rational frequencies in some interval. Thus we obtain both minimizing and minimax class-1 periodic orbits. If these minimax periodic orbits are elliptic, then oscillating orbits of class 2 and so forth can be obtained (recall Sec. II.C).

VII. QUASIPERIODIC ORBITS

In addition to the periodic orbits found in the previous section, there are also quasiperiodic orbits that minimize the action. In fact, we shall show that any rotational invariant circle is minimizing. Remarkably, however, minimizing quasiperiodic orbits exist for any twist map and any rotation frequency—not just for maps close enough to the integrable case and for Diophantine frequencies, as one might expect from the KAM theorem. When an invariant circle with a given frequency is destroyed, the corresponding minimizing quasiperiodic orbit can no longer densely cover a circle; in fact, it covers a Cantor set and is called a "cantorus." Cantori have an infinite set of gaps through which chaotic orbits can leak, and, as we shall see in Sec. IX, the leakage through cantori can be extremely slow.

We shall follow Aubry and Katok and obtain orbits with irrational ω by considering the limit of a set of m/n minimizing orbits as the period approaches infinity and the frequency approaches ω. This approach follows the ideas used in the numerical experiments of Greene, discussed in Sec. II.C. A more direct approach was developed by Mather (1982), who studied the action for curves introduced by Percival (1979a).

The analysis in this section is somewhat more formal than the rest of this article, and some of the proofs are omitted.

A. Circle maps

A rotational invariant circle can be described by a function $y=Y(x)$, which is periodic in x (Fig. 43). When restricted to the invariant circle the map becomes

$$(x',y')=T(x,Y(x)) \ . \tag{7.1}$$

A projection onto the x axis is denoted by the symbol π; thus for a point $z=(x,y)$

$$\pi(z)=x \ . \tag{7.2}$$

Equation (7.1) defines a function α, a "circle map," through

$$x'=\alpha(x)=\pi(T[x,Y(x)]) \ . \tag{7.3}$$

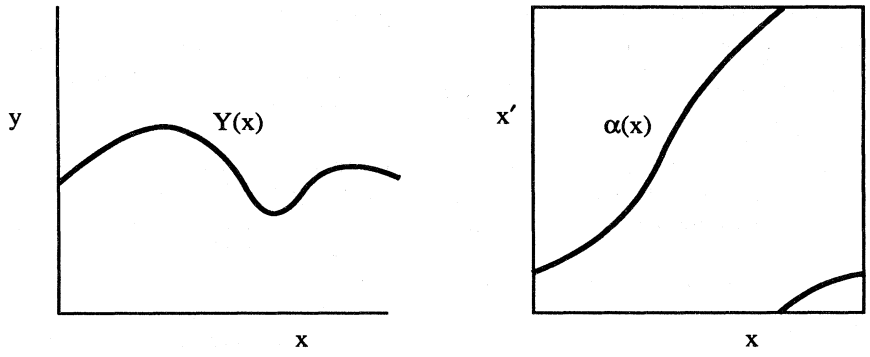

FIG. 43. Rotational invariant circle $Y(x)$ and corresponding circle map $\alpha(x)$. Here α is shown wrapped onto the torus $[0,1]\times[0,1]$.

Since the map is periodic with period one, we have $x'(x+1,y)=x'(x,y)+1$ and therefore

$$\alpha(x+1)=\alpha(x)+1 \ .$$

Thus α is a degree one circle map (see Appendix B). In fact, since T is a homeomorphism and Y is Lipschitz, the circle map α is a homeomorphism as well. A classic theorem of Poincaré implies that any homeomorphism of the circle has a unique rotation number (Appendix B), and thus all invariant circles of twist maps have one as well.

B. Invariant circles are minimizing

In any discussion of rotational invariant circles of twist maps, the concept of minimizing orbits arises naturally, since

Theorem. *Every orbit on a rotational invariant circle is minimizing.*

Proof. By Birkhoff's theorem (Sec. IV.C), every RIC is the graph of a Lipschitz function $Y(x)$. Let

$$S(x)=\int_{x_0}^{x} Y(\xi)\,d\xi \ , \tag{7.4}$$

integrating from some arbitrary point x_0. Define the function

$$H(x,x')=F(x,x')-S(x')+S(x) \ . \tag{7.5}$$

The derivatives of H are

$$\begin{aligned} H_1&=F_1(x,x')+Y(x) \ , \\ H_2&=F_2(x,x')-Y(x') \ . \end{aligned} \tag{7.6}$$

Following the discussion in Sec. V.A, each of these derivatives is zero exactly once, when $x'=\alpha(x)$. This implies that $H(x,\alpha(x))=H_0$ is constant and that all critical points of H occur on $x'=\alpha(x)$. Now Eq. (7.4) implies that $S(x+1)=S(x)+C$, where C is constant; so if F has zero net flux (5.7), then $H(x+1,x'+1)=H(x,x')$. Thus $H(x,x')$ satisfies the same growth condition (6.3) as $F(x,x')$, and H is bounded from below. Therefore

$$H(x,x')>H_0 \quad \text{for } x'\neq\alpha(x) \ . \tag{7.7}$$

Finally, suppose $\{x_j,\ldots,x_k\}$ is an orbit segment on the RIC and $\{\xi_j=x_j,\xi_{j+1},\ldots,\xi_{k-1},\xi_k=x_k\}$ is a deformation. Then the action of the deformed segment is

$$\begin{aligned} W\{\xi\}&=\sum_{i=j}^{k-1} H(\xi_i,\xi_{i+1})+S(x_k)-S(x_j) \\ &\geq (k-j)H_0+S(x_k)-S(x_j) \\ &\geq W\{x\} \ . \end{aligned} \tag{7.8}$$

Therefore the segment $\{x\}$ is minimizing.■

This theorem can be generalized in a limited sense to higher dimensions. The limitation is really the absence of a result comparable to Birkhoff's theorem: it is not known if every rotational invariant torus is a graph. However, upon assuming that the rotational tori are graphs and imposing the additional restriction that the torus must be a Lagrangian manifold (Herman, 1988), we have the

Theorem *(MacKay et al., 1989). For a symplectic mapping with uniformly positive-definite twist on $\mathbb{T}^N\times\mathbb{R}^N$ and zero net flux, every orbit on an invariant Lagrangian graph is minimizing.*

Proof. On a Lagrangian graph, $Y(x)=\nabla S(x)$. Use this S to define H as before; since the growth condition applies, we can follow Eqs. (7.4)–(7.8).■

C. Monotone sets

In order to show that the limit of a set of minimizing periodic orbits is a minimizing orbit, we need to use the fact that they are monotone [Eq. (6.17)]. In this section we discuss general properties of monotone sets before using these properties in the next section to prove the ex-

istence of quasiperiodic minimizing orbits.

As invariant set M is *monotone* if for all $z,\zeta\in M$,

$$\pi(z)<\pi(\zeta)\Longrightarrow\pi(T(z))<\pi(T(\zeta))\ , \tag{7.9}$$

where π is the projection (7.2). An orbit is monotone if the set formed from all its translates is monotone, i.e., (6.17). We showed in Sec. VI.C that Aubry's fundamental lemma implies that minimizing periodic orbits are monotone.

Monotone invariant sets for twist maps have nice properties:

Lemma. *A monotone invariant set M is a graph over x.*

Proof. Suppose not, then there are two points, $z=(x,y)$ and $\zeta=(\xi,\eta)$ in M which have the same x value: $x=\pi(z)=\xi=\pi(\zeta)$, but different y values, say $y>\eta$. However, the twist condition then implies that $\pi(T^{-1}z)<\pi(T^{-1}\zeta)$, which violates (7.9). Thus if $x=\xi$, then $y=\eta$, and $\pi(z)=\pi(\zeta)\Longrightarrow\pi(T(z))=\pi(T(\zeta))$■

Lemma. *Any limit of monotone orbits is monotone.*

Proof. Suppose that for each k, $\{x^{(k)}\}$ is a monotone orbit. Then points on the orbit must satisfy $x_i^{(k)}<x_j^{(k)}\Longrightarrow x_{i+1}^{(k)}<x_{j+1}^{(k)}$, and in the limit,

$$x_i^{(\infty)}<x_j^{(\infty)}\Longrightarrow x_{i+1}^{(\infty)}\le x_{j+1}^{(\infty)}\ . \tag{7.10}$$

The only possible problem is equality in (7.10). Suppose this occurs; then the twist condition implies that $x_{i+2}^{(\infty)}>x_{j+2}^{(\infty)}$ (Fig. 44). However, this implies there is a K such that for all $k>K$, $x_{i+2}^{(k)}>x_{j+2}^{(k)}$, contradicting the assumption that $\{x^{(k)}\}$ is monotone for all k. Thus the limit must be monotone.■

Lemma. *The closure of a monotone invariant set is monotone.*

Proof. Let $z_0=(x_0,y_0)$ and $\zeta_0=(\xi_0,\eta_0)$ be points in the closure of a monotone set M. Continuity of T and monotonicity of M implies that

$$x_0<\xi_0\Longrightarrow x_n\le\xi_n\ . \tag{7.11}$$

However, as in (7.10), equality is forbidden by the twist condition.■

Monotone states have a rotation number. This can be seen most easily by referring to well-known results on one-dimensional maps of the circle (Appendix B). First we show that the restriction of the twist mapping to a monotone set is equivalent to a mapping on the circle:

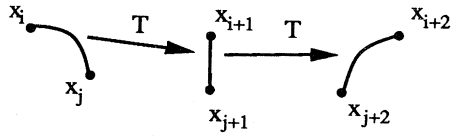

FIG. 44. Disallowing equality in Eq. (7.10). The twist condition implies that if two points are ever vertically aligned, then their order changes on the next iteration.

Lemma *(Katok, 1982). If T is a twist mapping and M is a monotone set, then the mapping from $\pi(M)$ to $\pi(T(M))$ can be extended to a homeomorphism $x'=\alpha(x)$ for $x\in\mathbb{R}$ satisfying $\alpha(x+1)=\alpha(x)+1$.*

Proof. The closure of M is monotone; so α can be extended to this by continuity. The complement of this closure is a disjoint union of open intervals. Extend α to these by linear interpolation for $x\in[0,1]$ and then continue to $\mathbb{R}$ by periodicity. Thus α is continuous, and because T is invertible, it has a continuous inverse.■

In Fig. 45, we sketch the construction of $\alpha(x)$ for a (2,5) monotone orbit. For an (m,n) orbit there are m inequivalent translations in the (x,y) plane. For the (2,5) case they are $\{x\}=\{\ldots,x_0,x_1,x_2,x_3,x_4,\ldots\}$ and $\{\xi\}=\{\ldots,x_3-1,x_4-1,x_0+1,x_1+1,x_2+1,\ldots\}$. We show part of the real line (of length 2) in Fig. 45, and the five points of each of the orbits $\{x\}$ and $\{\xi\}$ which lie in this segment. Define the function $\alpha(x)$ on the orbit so that $\alpha(x_t)=x_{t+1}$ and similarly $\alpha(\xi_t)=\xi_{t+1}$. Since the set of all translations is monotone, $\alpha(x)$ is a strictly increasing function. Thus defining $\alpha(x)$ by interpolation between the points x_t and ξ_t and between ξ_t and x_{t+1} gives a homeomorphism of the circle.

To see how this fails for a nonmonotone orbit, consider a configuration of type (2,4), Fig. 46. Recall from Sec. VI.D that if $x_2\neq x_0+1$, then this orbit cannot be minimizing because it cannot be monotone. [Indeed we showed that only orbits with (m,n) coprime can be monotone.] Though x_t increases with t, monotonicity fails because the translation $\xi_t=x_{t+2}-1$ is not well ordered with respect to x_t. In the figure we see that although $\xi_1>x_1,\xi_2<x_2$. This is reflected by a nonmonotonic segment in the induced $\alpha(x)$, which is therefore not a homeomorphism.

Theorem (B2) in Appendix B shows that every orbit of a homeomorphism of the circle has a rotation number, and the rotation number is the same for all orbits. So all monotone states have unique rotation numbers. Furthermore, the rotation number is a continuous function on monotone states:

Lemma. *The rotation number of the limit of a sequence of monotone states is the limit of the rotation numbers of the sequence.*

Proof. First we show that nearby monotone states have nearby rotation numbers. Let $\{x\}$ and $\{\xi\}$ be two monotone orbits, and suppose there are δ and T such that $|x_t-\xi_t|<\delta$ for all $0\le t\le T$. From Lemma (B1) in Appendix B it follows that there is a frequency $\omega(x)$ such that $|x_t-x_0-t\omega(x)|<1$, and similarly for ξ. Thus

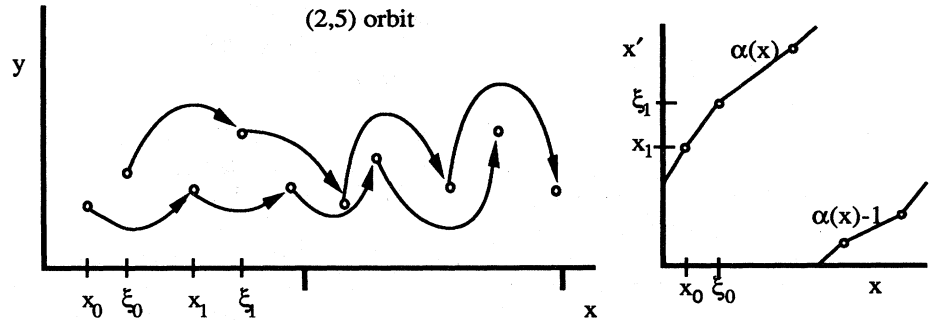

FIG. 45. Construction of the homeomorphism α for a (2,5) monotone orbit.

$$|x_t-\xi_t-(x_0-\xi_0)-t[\omega(x)-\omega(\xi)]|\le 2$$
$$\Rightarrow |\omega(x)-\omega(\xi)|\le 2\frac{(1+\delta)}{T}\ . \qquad (7.12)$$

Now consider a sequence of monotone states $\{x^{(k)}\}$, with periods $n^{(k)}\to\infty$ such that $m^{(k)}/n^{(k)}\to\omega$. If $\{x^{(k)}\}$ approach a limit $\{x^{(\infty)}\}$, there are δ and N such that $|x_t^{(k)}-x_t^{(\infty)}|<\delta$ for all $0\le t\le n^{(k)}$ and $k\ge N$. Since the periods go to infinity, (7.12) implies that the rotation number of $\{x^{(\infty)}\}$ is the same as the limit of the rotation numbers of the $\{x^{(k)}\}$.■

With these properties of monotone orbits, we can now prove the existence of monotone quasiperiodic orbits for every irrational ω.

D. Existence of quasiperiodic orbits

In Sec. VI.D we proved that there is a minimizing monotone state for every (m,n). We now show that this is true for all ω.

Theorem *(Mather, 1982; Aubry and Le Daeron, 1983). There exists a minimizing, monotone state for every ω.*

Proof. Consider a sequence of periodic minimizing states $\{x^{(k)}\}$ such that $m^{(k)}/n^{(k)}\to\omega$ as $k\to\infty$. By the lemmas in Sec. VII.C we conclude that $\{x^{(k)}\}\to\{x\}$ is a monotone state with frequency ω.

To show the limiting state is minimizing, consider a segment $\{\xi^{(k)}\}$ which is a deformation of $\{x^{(k)}\}$ with $\xi_i^{(k)}=x_i^{(k)}$ and $\xi_j^{(k)}=\xi_j^{(k)}$. Furthermore, $\{\xi^{(k)}\}\to\{\xi\}$. Let

$$\epsilon^{(k)}=\max(|x_t^{(k)}-x_t|,|\xi_t^{(k)}-\xi_t|)\ \text{ for } i\le t\le j\ . \qquad (7.13)$$

Since $F(x,x')$ is differentiable, there is a constant K, independent of k, such that

$$|W\{x_i^{(k)},\ldots,x_j^{(k)}\}-W\{x_i,\ldots,x_j\}|\le K(i-j)\epsilon^{(k)},$$
$$k>N$$

and similarly for $\{\xi\}$. Hence the action of the deformation $\{\xi\}$ minus that of $\{x\}$ obeys

$$W\{\xi_i,\ldots,\xi_j\}-W\{x_i,\ldots,x_j\}$$
$$\ge W\{\xi_i^{(k)},\ldots,\xi_j^{(k)}\}-W\{x_i^{(k)},\ldots,x_j^{(k)}\}$$
$$-2K(i-j)\epsilon^{(k)},\quad k>N\ .$$

Now because $\epsilon^{(k)}\to 0$ as $k\to\infty$, and each $\{x^{(k)}\}$ is

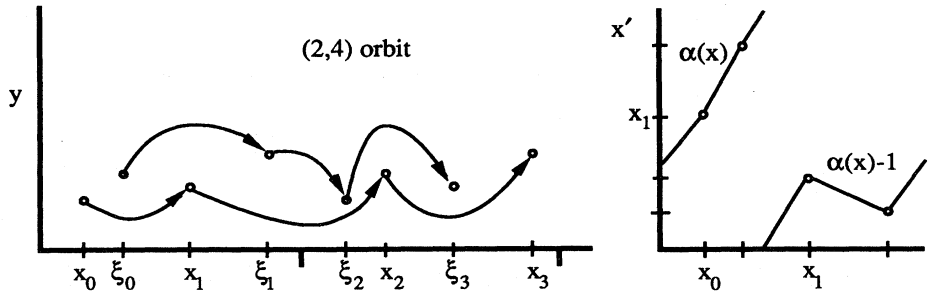

FIG. 46. Circle map construction for the (2,4) orbit. The map $\alpha(x)$ is not a homeomorphism because the orbit is not monotone.

minimizing, we have

$$W\{\xi_i,\ldots,\xi_j\}-W\{x_i,\ldots,x_j\}\geq 0\ ; \tag{7.14}$$

and so the limit is minimizing.■

A minimizing state obtained as a limit of periodic states is always recurrent, because there are periodic states arbitrarily close; thus such states are quasiperiodic. There are other minimizing states that are not recurrent. We shall discuss these below when we consider heteroclinic orbits.

If $\{x_t\}$ is a quasiperiodic orbit, then $\{x_{t+n}-m\}$ is another such state. These are never identical; otherwise the orbit would be periodic instead of quasiperiodic. Thus we have obtained a countable family of such states. This family is monotone, or totally ordered, by Aubry's fundamental lemma. In fact, the totality of minimizing states for a frequency ω is a closed monotone set (Aubry, 1983b).

The theorem showing that a limit of periodic states is quasiperiodic is of practical importance. For example, if one would like to study the properties of a particular quasiperiodic state, it is sufficient to study nearby periodic states and consider the limiting behavior of these properties. This was the approach pioneered by Greene in his studies of the breakup of invariant circles (Greene, 1979); recall Sec. II.C.

E. Cantori

We have seen that quasiperiodic minimizing orbits exist for all ω, for any twist mapping. Of course, this is not surprising for the case in which the mapping differs only slightly from an integrable mapping and the frequency satisfies a Diophantine condition: these orbits lie on the invariant circles of the KAM theorem (Sec. III.B). However, the KAM theorem applies only to this slightly perturbed case, while the Aubry-Mather theorem applies to any twist mapping. Furthermore, we have seen from Birkhoff's theorem that invariant circles typically do not exist when the nonlinear potential energy is sufficiently strong (Sec. IV.D). What do the minimizing quasiperiodic orbits become when there are no invariant circles? The answer is provided by the following

Theorem. *Let $\{x\}$ be a quasiperiodic minimizing orbit with frequency ω. The closure of $\{x\}$, $M_\omega\{x\}$ is either an invariant circle or an invariant Cantor set.*

Proof. Since the minimizing orbit is monotone, its closure can be extended to a homeomorphism, $\alpha(x)$ of the circle. Theorem (B3) in Appendix B implies that if the rotation number is irrational, the set of limit points of the orbit of any point is unique, invariant, and is either the entire circle or a Cantor set. Since we have assumed that the minimizing orbit is recurrent, then its closure is in fact this set of limit points of α.■

We remind the reader of the definition and some of the properties of Cantor sets in Appendix B.

Percival called the invariant Cantor sets "cantori"; we show a cantorus for the standard map in Fig. 47. He suggested the existence of cantori based on a variational principle for quasiperiodic orbits (Percival, 1979b); Aubry (1978) independently suggested their existence. Furthermore, Percival explicitly constructed cantori for a particular family: the sawtooth map (Percival, 1979b; Aubry, 1983a).

A cantorus is an invariant set that is "trying" to be a rotational invariant circle; however, orbits on this set fail to cover the circle: they never fall in a countable set of open intervals, or gaps—in fact, any Cantor set is equivalent to an interval with a countable family of deleted gaps. The end points of these gaps are quasiperiodic minimizing orbits, but the cantorus has uncountably many more orbits on it than these end points.

However, the structure of the cantorus is determined by the orbits of these end points, because the orbit of any point on the cantorus must densely cover the cantorus. We can "construct" a cantorus by imagining first that an invariant circle develops a single gap (imagine, if you will, that nearby islands on either side of the invariant circle have grown and squeezed a hole in the circle). Let x_0^l and x_0^r be the left and right end points of this gap. By definition of a gap there can be points in the interval (x_0^l,x_0^r) that are on a minimizing orbit with frequency ω. The orbits of each of the end points are distinct and quasiperiodic. Furthermore, because minimizing orbits are ordered, the iterates of these end points do not cross upon iteration; thus $x_t^l<x_t^r$. Since there are no minimiz-

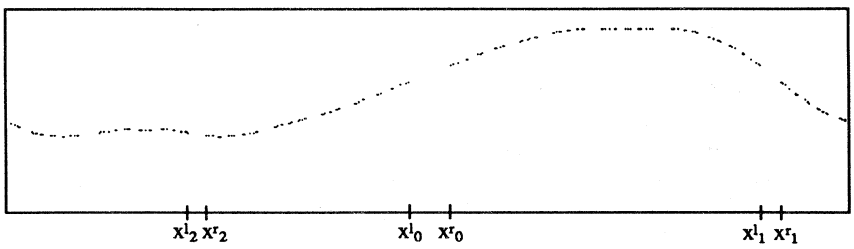

FIG. 47. Cantorus for $w=1/\gamma^2$ and $k=1.0$ for the standard map. The largest gap forms around $x=0$, the "potential minimum."

ing points in the interval (x_0^l, x_0^r), there can be none in the intervals (x_t^l, x_t^r). Thus each interval (x_t^l, x_t^r) is also a gap and each of these is distinct because the orbit is quasiperiodic. The total length of the gaps (in x) is at most 1; so the length of the iterate of any gap must eventually go to zero.

Thus we can speak of "iterating" a gap; however, what we really mean is that we iterate the end points of the gap—these are points on the cantorus itself. The iterates of a gap form a *family*. Since any Cantor set has at most a countable set of gaps, there are at most a countable set of families of gaps in a cantorus; typically we observe just one: every gap is the iterate of a single gap [though the example of Greene *et al.* (1987) probably has two families for some parameter values; see Ketoja and MacKay (1989)].

Cantori are typically hyperbolic, though I do not know of any theorem that guarantees this in general [when k is large enough all cantori of the standard map are hyperbolic; see Goroff (1985) and Veerman and Tangerman (1991)]. The hyperbolicity is measured by a Lyapunov multiplier, which is obtained from the linearized mapping along a segment of length n of the orbit: $|\mathrm{Tr}(M^n)|^{1/n} \to \lambda$, as $n \to \infty$. In numerical studies the Lyapunov multiplier is observed to grow continuously from 1 when a cantorus is formed.

When the Lyapunov multiplier is larger than 1, the iterate of any gap has a length that eventually must approach zero as λ^{-n}. This implies that the Hausdorff dimension of the cantorus is zero (Li and Bak, 1986; MacKay, 1987). This is remarkable, since it implies that when an invariant circle breaks, its length falls immediately to zero; furthermore, its dimension discontinuously changes from 1 to zero (providing it becomes hyperbolic).

F. Characterization of the set of minimizing orbits

So far we have shown that there exist minimizing orbits for each ω, and that these orbits are monotone. However, there could be other minimizing orbits that are not covered by these results. Here we mention some properties of the complete classification of the set of minimizing orbits (Aubry and Le Daeron, 1983; Mather, 1982, 1985).

Aubry's fundamental lemma implies that periodic minimizing orbits are monotone. This can be generalized to any minimizing orbit. Thus every minimizing orbit has a frequency ω, and for each ω the set M_ω of all minimizing orbits is monotone. Furthermore, if there is a rotational invariant circle with irrational frequency ω, then every minimizing orbit of frequency ω is recurrent.

We have seen that to every minimizing quasiperiodic orbit there corresponds a homeomorphism of the circle; however, it is not obvious that different minimizing orbits correspond to the same homeomorphism. One could imagine that the closures of different orbits might give rise to disjoint invariant circles, or disjoint Cantor sets. However, when the twist is monotone this cannot happen.

In addition to the periodic and quasiperiodic minimizing orbits, a new class, the nonrecurrent orbits, must be considered. Since the set of minimizing orbits of frequency ω is monotone, the nonrecurrent minimizing orbits must lie in the gaps of the recurrent minimizing orbits.

When $\omega = m/n$ the nonrecurrent orbits are crossing points of the stable and unstable manifolds of the minimizing orbit: they are homoclinic to the minimizing (m,n) orbit. There are two such orbits that are minimizing. One is the "advancing" homoclinic orbit. As $t \to -\infty$, this orbit is asymptotic to the left end point of a gap, while when $t \to \infty$, it is asymptotic to the right end point. The other orbit is the "retreating" homoclinic orbit. These orbits lie on the upper and lower separatrix of the resonance, and we shall discuss them in more detail in Sec. VIII.

When ω is irrational the nonrecurrent minimizing orbit lies in the gaps of the cantorus; since the gap widths must shrink to zero, it is homoclinic to the cantorus.

G. Mather's ΔW

The nonexistence of an invariant circle is implied by the existence of a nonminimizing orbit with frequency ω. In particular, if the limit of the minimax periodic orbits as $m/n \to \omega$ is an orbit with larger action than the minimizing quasiperiodic orbit, then there is no invariant circle.

Theorem *(Mather, 1986). Let* $\{x^{(k)}\}$ *and* $\{\xi^{(k)}\}$ *be sequences of minimizing and minimax* (m_k, n_k) *periodic states, respectively, such that* $m_k/n_k \to \omega$. *Then the limit of action differences*

$$\Delta W_\omega = \lim_{k\to\infty} [W_{(m_k,n_k)}\{\xi^{(k)}\} - W_{(m_k,n_k)}\{x^{(k)}\}] \qquad (7.15)$$

exists and is non-negative. If $\Delta W_\omega > 0$, *there is no invariant circle with frequency* ω.

We shall see in Sec. VIII.C that the quantity ΔW_ω can be interpreted as the flux through the minimizing set. It is therefore natural that $\Delta W_\omega = 0$ when there is an invariant circle.

VIII. FLUX

Flux is the area per unit time that crosses from one side of a surface in phase space to another; we defined it in Sec. II.D. A calculation of flux can be used to obtain estimates of transport rates, for example, the transition time for trajectories to move from one region of phase space to another. Suppose we consider trajectories starting in the region $y < y_0$ and would like to estimate the time to enter the region $y > y_1$. If there is an invariant circle in the annulus $y_0 < y < y_1$, then of course this time

would be infinite. More generally, the transit time would be long if there were rotational circles with small flux; the minimum flux rotational circle would be most restrictive.

Wigner (1937) proposed that finding the minimum flux surface for a Hamiltonian representing a chemical system would yield good estimates of reaction rates. His formulation was variational. In this section we shall use a different variational principle, the one for twist maps, and present evidence that minimum flux curves are associated with noble cantori. First we discuss techniques for computing flux.

A. Partial barriers and turnstiles

We call a curve that has small flux a *partial barrier*: chaotic orbits leak through the barrier, but they do so slowly. In this section we discuss several ways of constructing rotational partial barriers; each way uses minimizing orbits (MacKay *et al.*, 1984). One reason for using these is that for the integrable case a minimizing orbit lies on a rotational invariant circle; and since minimizing orbits are monotone, one might expect them to approximate such a circle even for the case in which there are none.

Furthermore, the monotonicity of these orbits allows a simple construction of a rotational (noninvariant) circle. To do this we use the notion of *gaps*. Let x_0 be a point on a monotone orbit; then the gap g_0 is the segment between x_0 and the nearest neighbor to the right on the orbit of x_0 or any of its translates (Fig. 48). Monotonicity implies that upon iteration $g_0 \to g_1$ becomes a gap between x_1 and its nearest neighbor on the right. We call the set of iterates of a gap a *family* of gaps; a monotone orbit can have more than one family of gaps, though there can be at most a countable number.

1. Periodic orbits

One way to construct a rotational partial barrier is to use the minimizing and minimax (m,n) orbits. Choose any gap in the minimizing orbit, call it the *principal gap*, and fill it with an arbitrary curve, $\mathcal{L}_0$, which also goes through the minimax orbit (see Fig. 48). The remaining gaps are filled with the $n-1$ preimages of this curve, $\mathcal{L}_{-1},\mathcal{L}_{-2},\ldots,\mathcal{L}_{-n+1}$. The resulting curve is a rotational circle, a "partial barrier" connecting all the points on the two (m,n) orbits.

This curve defines a partial barrier that divides the cylinder. To move from one side of the partial barrier to the other, a trajectory must cross; it can do so because the partial barrier is typically not an invariant curve. In fact, when iterated, each $\mathcal{L}_t \to \mathcal{L}_{t+1}$, which is another segment of the partial barrier except that $\mathcal{L}_0 \to \mathcal{L}_1$, which is not part of the barrier. To visualize the flux through the partial barrier, take the preimage of $\mathcal{L}_{-n+1}$ to obtain a second curve, $\mathcal{L}_{-n}$, in the principal gap, the dashed curve in Fig. 49. It must connect the end points of g_0, because the end points lie on a periodic orbit.

Using the partial barrier, i.e., the solid curve in Fig. 49, to provide a definition of "below" and "above," we see that the only region that crosses from below to above on one iteration of the mapping is the region below $\mathcal{L}_0$ and above $\mathcal{L}_{-n}$. Similarly, the region below $\mathcal{L}_{-n}$ but above $\mathcal{L}_0$ crosses from above to below upon one iteration. These areas define the upward and downward fluxes through the barrier. Because the net flux is zero, the fluxes up and down are equal; therefore $\mathcal{L}_{-n}$ and $\mathcal{L}_0$ must cross at least once, giving the characteristic figure-eight structure shown in Fig. 49, which we call a *turnstile* (MacKay *et al.*, 1984). This is because it acts as a "rotating door," dumping all the area in its left lobe above, and all the area in its right lobe below the partial barrier each iteration.

As we shall see below, the flux is independent of the construction of the partial barrier, providing it connects neighboring points on the minimizing orbit and goes through the minimax orbit. Thus the arbitrariness in the choice of $\mathcal{L}_0$ is not important, and we can think of the minimizing-minimax pair of orbits themselves as defining a partial barrier.

Turnstiles can be more complicated than we indicated in Fig. 49. For example, there is nothing that prevents the turnstile from looking like Fig. 50. In this case the flux is the shaded region shown. Though we have never seen a turnstile with this structure, it could occur in physical examples.

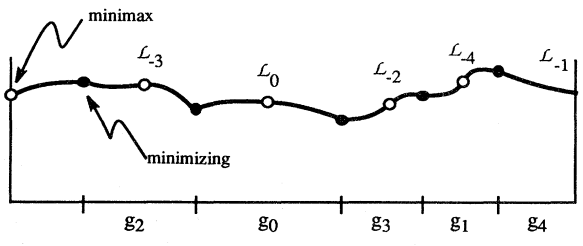

FIG. 48. Partial barrier for the (2,5) orbits. The orbit of gaps, g_t, is shown along the x axis. $\mathcal{L}_{-t}$ are the preimages of an initial segment $\mathcal{L}_0$ of arbitrary shape in the principal gap of the minimizing orbit.

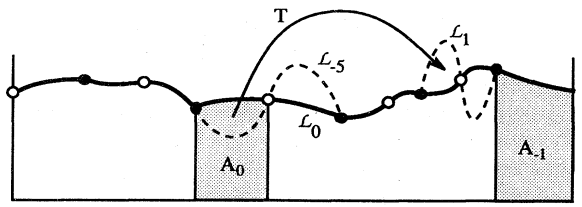

FIG. 49. Turnstile for and area under periodic orbits. Turnstile in the gap g_0 gives the flux crossing the partial barrier. The areas A_0 and A_{-1} are used to obtain Eq. (8.4).

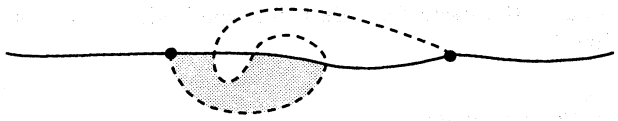

FIG. 50. Possible turnstile shape.

2. Homoclinic orbits

Probably the most familiar case of a turnstile occurs in the separatrix of a hyperbolic orbit. The hyperbolic minimizing orbit has stable and unstable manifolds, W^s and W^u, and, as discussed in Sec. VII.F, advancing and retreating minimizing homoclinic orbits. For definiteness, consider the (0,1) orbits and the advancing homoclinic orbit. A partial barrier is formed by choosing a gap g_0 (the principal gap) in the homoclinic orbit and closing it with a segment $\mathcal{U}_0$ of W^u. Preimages of $\mathcal{U}_0$ converge to the minimizing (0,1) orbit. For positive t, $\mathcal{U}_t$ oscillates increasingly wildly; so in order to construct a well-behaved rotational circle we switch to segments of the stable manifold $\mathcal{S}_t$. These converge to the (0,1) orbit as $t \to \infty$. In this way we construct a piecewise smooth rotational partial barrier, with a discontinuity in slope at the right end point of g_0, shown as the solid curve in Fig. 51. As before, the turnstile is obtained by taking the preimage of the partial barrier. Each segment has a preimage on the partial barrier, except for $\mathcal{S}_1$, which becomes $\mathcal{S}_0$ in the principal gap, and gives the turnstile. Since the advancing minimax orbit lives in the gaps of the advancing minimizing orbit and is homoclinic to the (0,1) orbit, a point on the minimax orbit must be on $\mathcal{U}_0$, as shown in Fig. 51; therefore the turnstile must have at least two lobes. Just as in the periodic orbit construction, the two lobes of the turnstile correspond to areas that cross the partial barrier each iteration.

The construction for the (m,n) case is similar. One gap in the (m,n) orbit is closed with segments of unstable and stable manifolds as before; the switch occurs at an arbitrary point on the advancing minimizing homoclinic orbit (labeled m_0 in Fig. 52). Taking $n-1$ preimages gives curves that fill in the remaining gaps of the (m,n) orbit. Thus there is a discontinuity in the slope of the

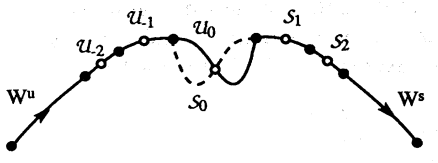

FIG. 51. Partial barrier for an advancing minimizing orbit homoclinic to the (0,1) minimizing orbit. It consists of a segment of unstable manifold from the left point on the (0,1) orbit to some arbitrarily chosen point on the advancing minimizing homoclinic orbit. From there we switch to stable manifold leading to the right point on the (0,1) orbit.

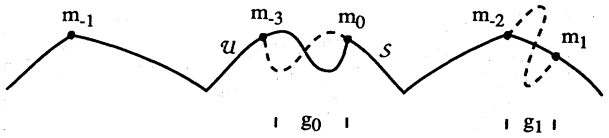

FIG. 52. Upper partial barrier for the (1,3) orbit. The turnstile, dashed and solid curves in the gap g_0, iterates to the curves in g_1. Only the left lobe of the turnstile moves from below to above the partial barrier in one iteration.

partial barrier at the points m_t for $-n < t \le 0$. The flux through the partial barrier is localized to one turnstile, that contained in the gap g_0 between m_{-n} and m_0.

3. Resonances

The construction of a partial barrier for the advancing and retreating homoclinic orbits of an (m,n) orbit leads to a precise definition of a resonance: it is the area contained between these upper and lower partial barriers. The upper turnstile area of an (m,n) resonance gives the area that makes a transition from inside the m/n resonance to some resonance above (m,n). Similarly, the lower turnstile represents the area making a transition to below (m,n).

The shape of the resonance depends on the choice of homoclinic point at which we switch from unstable to stable segments. However, the turnstile area is independent of this choice, since any iterate of the turnstile must have the same area as the original one. Similarly, the total area of the resonance is independent of the choice of homoclinic point, because for a different choice the shape of the resonance changes by the addition of one entering and the deletion of one exiting turnstile area. These are equal since the net flux is zero.

4. Cantori

A similar partial barrier can be constructed for a cantorus. Choose a gap g_0 in the cantorus. Since the cantorus lies on a Lipschitz graph and is monotone, the length of any gap must go to zero far enough in the future and in the past. The stable manifold theorem, Sec. II.B, implies that there are manifolds $\mathcal{S}_t$ and $\mathcal{U}_t$ that connect the end points of g_t and that approach the cantorus as $t \to \infty$ and $t \to -\infty$, respectively. A partial barrier is obtained by forming the curve from $\mathcal{U}_t$ for $t \le 0$ and $\mathcal{S}_t$ for $t > 0$ (see Fig. 53). If there is only one family of gaps, then the resulting curve will be a rotational circle and will form the partial barrier; otherwise, since there is a countable number of gaps, we can repeat the construction for each family. The preimage of $\mathcal{S}_1$ lies in g_0 but would coincide with $\mathcal{U}_0$ only if the barrier we constructed had been an invariant circle, contrary to assumption. The segments $\mathcal{S}_0$ and $\mathcal{U}_0$ must cross at least once, since they necessarily go through the minimax orbit (which

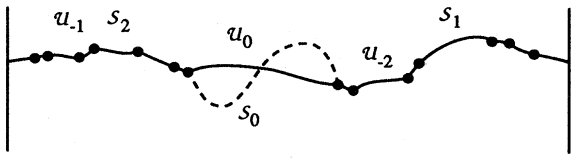

FIG. 53. Partial barrier in a cantorus.

lives in the gaps and is homoclinic to the cantorus). The combination of $\mathcal{S}_0$ and $\mathcal{U}_0$ forms a turnstile. The flux through the cantorus is the area in one lobe of the turnstile, as before. Note that even though there are an infinite number of gaps in the cantorus, the entire flux is localized to the principal gap by this construction.

B. Areas and actions

1. Fundamental formula

Areas of resonances and of turnstiles are both needed for the theory of transport. An obvious way to calculate them is to approximate the boundaries by closely spaced points and then to use numerical integration; however, this is not the best way. In fact, these areas can be obtained solely from the action of the minimizing and minimax orbits making up the partial barriers (Bensimon and Kadanoff, 1984; MacKay *et al.*, 1984).

The basic formula relating area to action is given by Eq. (5.6). In fact, this is a relation between algebraic area and action. As in Sec. V.A, let $\mathcal{C}$ be a directed curve in the phase plane; we parametrize it by $\lambda\in[0,1]$, so that

$$\mathcal{C}(\lambda)=\{x(\lambda),y(\lambda)\} . \tag{8.1}$$

Let A be the *algebraic* area "under" $\mathcal{C}$, i.e., the value of Eq. (4.1). For the simple situation depicted in Fig. 39, A is merely the geometric area. If, however, $\mathcal{C}$ intersects itself or if $y(\lambda)$ is negative for some range of λ, then the sign of the areas of these regions will change, and A will not be the geometric area between $\mathcal{C}$ and the x axis. In any case we shall still refer to A as the area "under" $\mathcal{C}$, though some regions may be included with negative sign. The image of $\mathcal{C}$ is denoted $\mathcal{C}'$ and has an area A'.

Let $F(x,x')$ be the generating function of the twist map T from the initial point with configuration $x(\lambda)$ to its image $x'(\lambda)$. By Eq. (5.1)

$$\begin{aligned}\frac{dF}{d\lambda}&=F_1\frac{dx}{d\lambda}+F_2\frac{dx'}{d\lambda}\\&=y'\frac{dx'}{d\lambda}-y\frac{dx}{d\lambda} .\end{aligned} \tag{8.2}$$

Integrating both side with respect to λ, we obtain

$$\begin{aligned}\Delta F&\equiv F[x(1),x'(1)]-F[x(0),x'(0)]\\&=A'-A .\end{aligned} \tag{8.3}$$

This is the basic formula from which all the others follow.[4]

2. Periodic orbits

The flux through the turnstile in a pair of (m,n) periodic orbits is easily obtained from the fundamental formula. Let $\{m_t\}$ denote the minimizing orbit, and $\{s_t\}$ the minimax orbit. Let A_0 be the area under the portion of the $\mathcal{L}_0$ connecting m_0 to s_0 (Fig. 49). Similarly, A_t represents the area under the portion of the iterate $\mathcal{L}_t$ connecting m_t to s_t. The fundamental formula (8.3) implies that

$$A_t-A_{t-1}=F(s_{t-1},s_t)-F(m_{t-1},m_t) . \tag{8.4}$$

The area of turnstile, A_0-A_{-n}, is obtained by adding successive iterates of (8.4):

$$\begin{aligned}\mathcal{F}=A_0-A_{-n}&=\sum_{t=-n+1}^{0}[F(s_{t-1},s_t)-F(m_{t-1},m_t)]\\&=\sum_{t=0}^{n-1}[F(s_t,s_{t+1})-F(m_t,m_{t+1})]\\&=W_{(m,n)}\{s\}-W_{(m,n)}\{m\}\equiv\Delta W_{(m,n)} .\end{aligned} \tag{8.5}$$

Thus the flux is simply the difference in action between the minimax and minimizing orbits. It therefore does not depend on the choice of $\mathcal{L}_0$, or indeed in which gap the turnstile is placed.

3. Stable and unstable segments

The formulas for the flux through homoclinic orbits and cantori also follow from Eq. (8.3), but we cannot rely on periodicity, as in Eq. (8.5). Instead we use the fact that in both cases the gaps shrink to zero in the past and in the future.

Two points in phase space, z_0 and w_0, are called *future asymptotic* if they are distinct, but their orbits approach each other asymptotically, so as to become indistinguishable at sufficiently long times in the future:

$$\lim_{t\to\infty}|z_t-w_t|=0 \tag{8.6}$$

where $||$ represents any norm. Similarly, they are *past asymptotic* if they are distinct and their orbits approach each other asymptotically in the past:

$$\lim_{t\to-\infty}|z_t-w_t|=0 . \tag{8.7}$$

Points that are both future and past asymptotic are

[4]This relation, and the others that follow, can be generalized not only to maps that do not satisfy the twist condition, but also to those that are not area preserving (Easton, 1991).

homoclinic (to each other). If v_0 is past asymptotic to z_0 and future asymptotic to z_0, then it is *heteroclinic* from z_0 to w_0.

If an orbit z_t is hyperbolic, then the set of points that are future or past asymptotic to z_0 forms two smooth curves without self-intersection, crossing transversely at z_0, called the stable and unstable manifolds of z_0 (recall Sec. II.B). All points on the same stable manifold are future asymptotic, and all points on the same unstable manifold are past asymptotic. Given two such points, we call the piece of invariant manifold between them a stable or unstable *segment.* As we have seen, partial barriers for cantori and minimizing homoclinic orbits are made from such segments.

We can find stable (unstable) segments numerically by taking the limit of backward (forward) iterates of straight lines joining corresponding points of two future (past) asymptotic orbits. Thus if z_0 and w_0 are future asymptotic, let $\mathcal{L}_j, j>0$, be the directed straight-line segment from z_j to w_j (Fig. 54). Then the stable segment joining z_0 to w_0 is

$$\mathcal{S}_0=\lim_{j\to\infty} T^{-j}(\mathcal{L}_j) \ . \tag{8.8}$$

Similarly, a pair of past asymptotic points gives an unstable segment

$$\mathcal{U}_0=\lim_{j\to\infty} T^{j}(\mathcal{L}_j) \ . \tag{8.9}$$

The images of a stable (unstable) segment are also stable (unstable) segments and are denoted $\mathcal{S}_t(\mathcal{U}_t)$.

Using the fundamental formula, the area below a stable or unstable segment can be expressed in terms of sums of action differences. Let $\{w_t\}$ and $\{z_t\}$ be a future asymptotic pair, and denote the action difference by

$$\Delta F_t \equiv F(w_t,w_{t+1})-F(z,z_{t+1}) \ . \tag{8.10}$$

Suppose $\mathcal{S}_t$ is a stable segment connecting the tth points on these orbits. We can parametrize it with λ, just as in Eq. (8.1), so that $\mathcal{S}_t(0)$ is z_t and $\mathcal{S}_t(1)$ is w_t. The area under $\mathcal{S}_t$, denoted A_t^s in Fig. 55, is obtained by iterating the fundamental formula (8.3):

$$\begin{aligned} A_t^s &= A_{t+1}^s-\Delta F_t \\ &= A_{t+k}^s-\sum_{j=0}^{k-1}\Delta F_{t+j} \ . \end{aligned} \tag{8.11}$$

Now because the action difference is taken between two

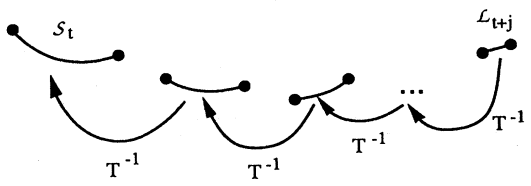

FIG. 54. Construction of the stable segment, following Eq. (8.8).

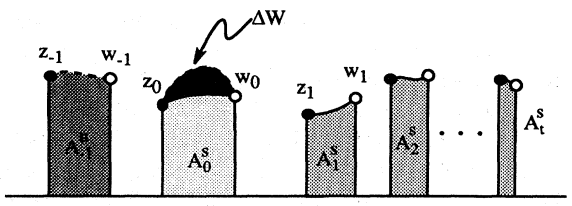

FIG. 55. Area under a future asymptotic pair. The stable segment connecting z_0 and w_0 has area A_0^s. The unstable segments are dashed. The action difference $\Delta W=A_0^u-A_0^s$ is the dark region, Eq. (8.13).

future asymptotic orbits, $A_{t+k}^s\to 0$ as $k\to\infty$; so the sum in (8.11) can be extended to ∞, yielding

$$A_t^s=-\sum_{j=0}^{\infty}\Delta F_{t+j} \ . \tag{8.12}$$

Since F is continuously differentiable, the convergence of the sum (8.12) is guaranteed if the union of the two orbits is monotone, since the sum of the gap widths is bounded by 1, or if the orbits are hyperbolic, since the points converge exponentially.

If $\{w_t\}$ and $\{z_t\}$ are past asymptotic, and A_t^u is the area under their unstable segment $\mathcal{U}_t$, then a similar calculation gives

$$A_t^u=\sum_{j=-\infty}^{-1}\Delta F_{t+j} \ . \tag{8.13}$$

Note that the $t=0$ term is not included here, and the sign is indeed different from the previous one.

C. Flux formulas

1. Homoclinic pair

We can combine the future (8.12) and past (8.13) sums if $\{z_t\}$ and $\{w_t\}$ are homoclinic. The signed area between the unstable and stable segments (positive where $\mathcal{U}_t$ is above $\mathcal{S}_t$) is given by

$$A_t^u-A_t^s=\sum_{j=-\infty}^{\infty}\Delta F_j \ . \tag{8.14}$$

By a slight abuse of notation, we can write (8.14) as a difference between the actions of the two orbits:

$$A_t^u-A_t^s=W\{w_t\}-W\{z_t\}=\Delta W \ . \tag{8.15}$$

This area, the dark region in Fig. 55, is the flux through a homoclinic pair of orbits. Thus we have shown that the flux is Mather's ΔW, (7.15).

Note that (8.15) is independent of t by area preservation—the region contained between the stable and unstable segments has the same area for all time. For example, we can let $\{z_t\}$ be a minimizing orbit $\{m_t\}$ and $\{w_t\}$ be the corresponding minimax orbit $\{s_t\}$, corre-

sponding to a separatrix or a cantorus. In this case the upward flux flowing between $\{m_t\}$ and $\{s_t\}$ is the difference in actions of these orbits, and (8.14) is the area of the left lobe of the turnstile in Fig. 52 or 53.

A more general application of (8.12) and (8.13) is to find the area of a region bounded by a set of points connected by an alternating sequence of stable and unstable manifolds (Easton, 1991). For example, suppose there are four points u_0, v_0, w_0, and z_0, and that the orbits $\{u_t\}$ and $\{v_t\}$ are future asymptotic, $\{v_t\}$ and $\{w_t\}$ are past asymptotic, and $\{w_t\}$ and $\{z_t\}$ are future asymptotic. Finally, the region is closed with an unstable manifold connecting z_0 with u_0; so $\{z_t\}$ and $\{u_t\}$ are past asymptotic. The area of the region is obtained by applying (8.12) or (8.13) to each segment to obtain

$$A = W\{u_t\} - W\{v_t\} + W\{w_t\} - W\{z_t\} \ . \tag{8.16}$$

We shall find application of (8.16) in Sec. IX.C.

2. Flux Farey tree

To actually compute the flux, one needs to use finite orbits, and it is often convenient to use periodic orbits (8.5). To be systematic we compute the flux for each rational on the Farey tree (recall Fig. 26). In Fig. 56 we show eight levels of the Farey tree beginning with the neighboring rationals $\frac{1}{3}$ and $\frac{1}{2}$ for the standard map. The ordinate is the log of the flux through the orbit for $k = k_{\rm cr}(\gamma)$, (2.23), and the abscissa is the frequency. This figure leads to several observations and conjectures about

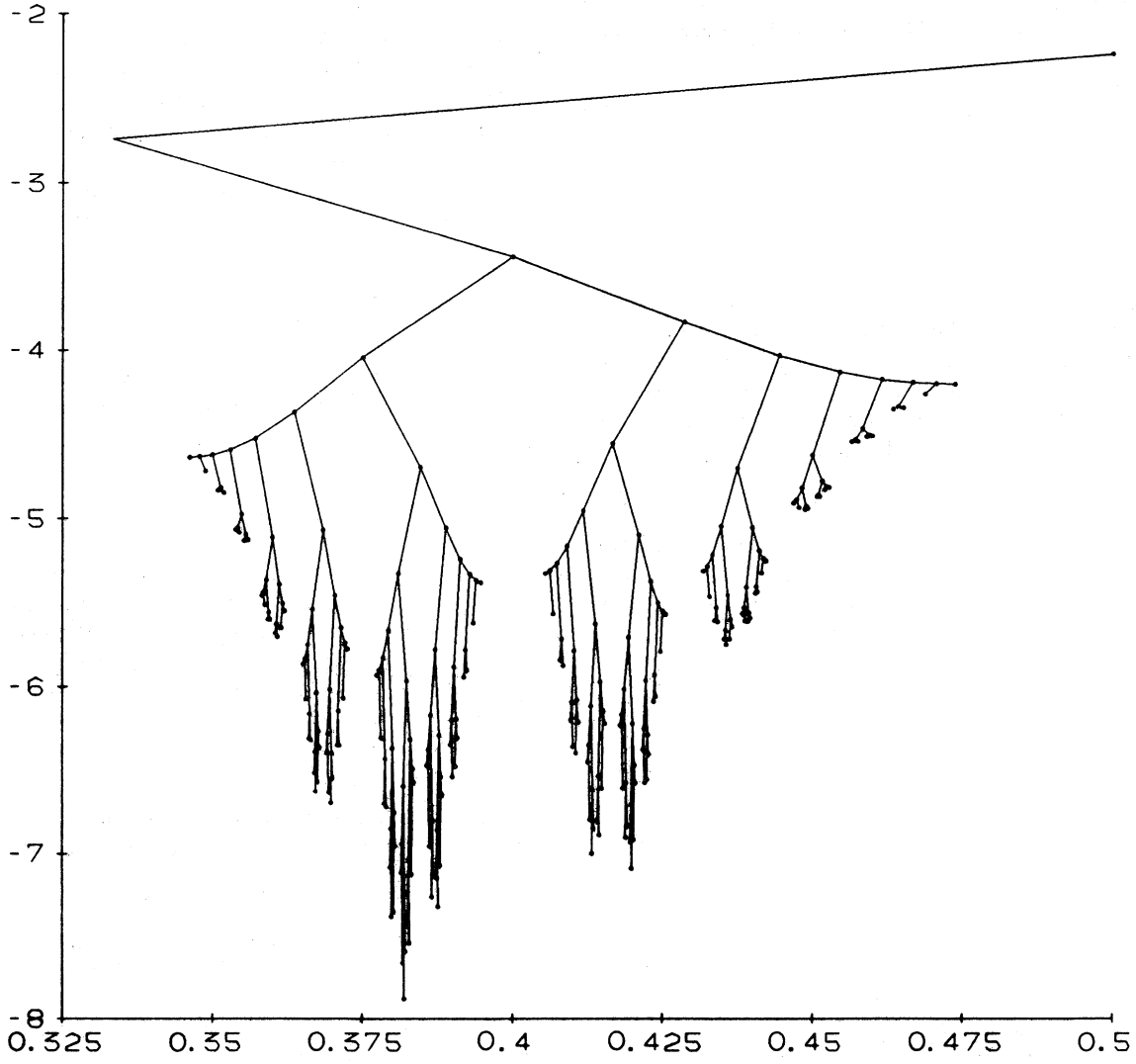

FIG. 56. Flux Farey tree for the standard map at $k = k_{\rm cr}(\gamma)$. Shown are eight levels of the tree beginning with the neighbors (1,3) and (1,2).

flux.

Notice that the flux through a daughter rational is never larger than the flux through either of its parents. We observe this to be true for any k, and indeed for any of the maps we have studied. For example, the flux through the (2,5) orbit is smaller than either of its parents (1,3) and (1,2). Though this property could be violated if the map had large Fourier coefficients at some frequencies, we conjecture that it is a general property of smooth maps for large enough levels on the tree.

Since irrationals are limits of infinite Farey paths, the flux through a quasiperiodic orbit is smaller than that through any of the rationals above it on the Farey tree. In this sense then, the cantori provide curves of local minimum flux. They form the most important barriers in any frequency interval.

A second observation from Fig. 56 is that, of the two daughters of a given rational, one always has smaller flux than the other. This rational corresponds to the Farey path that changes direction. Thus of the two daughters (3,8) and (3,7) of (2,5), the first has smaller flux. Recall from Fig. 27 that alternating directions on the tree correspond to a continued-fraction expansion whose elements are 1's. Thus the cantorus with the smallest flux below a rational (m,n) is the noble irrational corresponding to appending an infinite sequence of 1's to the continued fraction of (m,n). Again we conjecture that this is a general property of smooth maps for large enough levels on the tree.

For the special case of the standard map, this property appears to hold for all levels on the tree. For example, in the interval $[\frac{1}{3},\frac{1}{2}]$ the most noble irrational is $1/\gamma^2$. The corresponding Farey sequence, m_j/n_j, is

$$\frac{1}{2},\ \frac{1}{3},\ \frac{2}{5},\ \frac{3}{8},\ \frac{5}{13},\ \frac{8}{21},\ \frac{13}{34},\ \frac{21}{55},\ \frac{34}{89},\ \frac{55}{144}\cdots . \tag{8.17}$$

In Fig. 56 the lowest flux corresponds to the periodic orbit (55,144). The flux through the golden cantorus itself would be zero in Fig. 56, since $k=k_{\rm cr}(\gamma)=k_{\rm cr}(1/\gamma^2)$. In general, for the standard map, the golden cantorus has the smallest flux of any cantorus.

In fact, though it is difficult to see in Fig. 56, the flux through the orbits in the sequence (8.17) decreases geometrically with level (MacKay *et al.*, 1984),

$$\Delta W_{(m_j,n_j)}\sim C\xi^{-j},\quad \xi\approx 4.339\ . \tag{8.18}$$

The constant ξ can be computed using the renormalization analysis. This equation applies not just to the critical golden circle, but to any critical noble. A slight generalization applies to any "boundary circle" (Greene *et al.*, 1986; we shall discuss boundary circles in Sec. IX.C).

At $k=k_{\rm cr}(\gamma)$ there is only one invariant circle; so the flux through any sequence other than (8.17) converges to a nonzero constant. The convergence to a cantorus is faster than geometric. In fact, the convergence is related to the Lyapunov multiplier, λ, of the limiting orbit:

$$\Delta W_{(m_j,n_j)}\sim \Delta W_\omega + C\lambda^{-n_j}\ . \tag{8.19}$$

Equation (8.19) converges rapidly because n_j grows geometrically with level for an alternating path. This implies that an accurate calculation of ΔW for a highly unstable cantorus ($\lambda\gg 1$) can be made with a relatively short periodic orbit. Other properties of the cantori also converge with the same rapidity (MacKay *et al.*, 1984).

Every Farey path that eventually moves in one direction (either L or R) converges to a rational [from above or below, respectively; recall (3.12) and (3.13)]. The corresponding sequence of orbits converges either to the upper or lower homoclinic orbits. In Fig. 56 one can see that these sequences have well-defined nonzero limits for ΔW. These correspond to the fluxes through the upper and lower separatrices of the resonance and have different values in general. The separatrix fluxes are always larger than those through nearby cantori.

D. Area formulas

1. Cantorus area

To find the area under the partial barrier formed from a cantorus, suppose the cantorus has a single family of gaps, and let $\{z_t\}=\{x_t^l\}$ be the orbit of the left end points of a gap in the cantorus and $\{w_t\}=\{x_t^r\}$ be the orbit of the right end points. Backward iterates of the unstable segment of a gap and forward iterates of the stable segment form the cantorus partial barrier (recall Fig. 53). The area under a single segment of the partial barrier is given by either (8.12) or (8.13). To find the total area we simply sum over t. The area under all the segments $\mathcal{S}_t$ after time t is

$$\sum_{k=1}^{\infty} A_{t+k}^s = -\sum_{k=1}^{\infty}\sum_{j=0}^{\infty}\Delta F_{t+j+k} = -\sum_{k=1}^{\infty} k\Delta F_{t+k}\ . \tag{8.20}$$

This converges for hyperbolic orbits, since ΔF_t approaches zero exponentially. Similarly, Eq. (8.13) gives area under all the unstable segments for t or less as

$$\sum_{k=-\infty}^{0} A_{t+k}^u = -\sum_{k=-\infty}^{0} k\Delta F_{t+k}\ . \tag{8.21}$$

The total area under all the stable and unstable segments in the sum of (8.20) and (8.21):

$$A_\omega = -\sum_{t=-\infty}^{\infty} t[F(x_t^l,x_{t+1}^l)-F(x_t^r,x_{t+1}^r)]\ . \tag{8.22}$$

Suppose that there is only one family of gaps; then (8.22) includes all the gaps. Furthermore, when the cantorus is hyperbolic, it has zero length (indeed, zero dimension; recall Sec. VII.E). This implies that the area under the Cantor set itself is zero. Thus (8.22) is the area under the

cantorus partial barrier. Note that the area under a partial barrier is independent of the construction of the partial barrier itself, depending only on the orbits of the gap end points; thus we can refer to Eq. (8.22) as the area "under the cantorus."

2. Resonance area

Now we obtain the area under an upper partial separatrix for the simplest case of the (0,1) resonance. Let x_F denote the minimizing fixed point. Choose z_t to be x_F and w_t to be any point m_t^+ on the upper minimizing homoclinic orbit. The area under the unstable segment connecting x_F to m_t^+ is given by (8.12), and the area under the stable segment connecting m_t^+ to x_F is given by (8.13). Thus the total area under the upper separatrix is

$$A^+_{(0,1)}=\sum_{t=-\infty}^{\infty}[F(m_t^+,m_{t+1}^+)-F(x_F,x_F)]\ .$$

For the lower partial separatrix the unstable segment connects a point on the lower minimizing homoclinic orbit to x_F; so we define $z_t=m_t^-$ and $w_t=x_F$ and obtain

$$A^-_{(0,1)}=-\sum_{t=-\infty}^{\infty}[F(m_t^-,m_{t+1}^-)-F(x_F,x_F)]\ .$$

The change in sign arises from the reversed ordering of the points.

The analysis for an arbitrary (m,n) resonance is similar. Letting x_t represent the minimizing (m,n) orbit, choose a point m_t^+ on the upper homoclinic orbit in the gap to the right of x_t. As in Fig. 52, the upper boundary of the resonance in this gap is formed from the unstable manifold connecting x_t to m_t^+ and the stable manifold connecting m_t^+ to the right neighbor of x_t; the area under these segments is given by adding (8.12) and (8.13) as usual. Since this area is independent of the choice of gap, and there are n gaps in the (m,n) orbit, the area under the complete upper partial separatrix is simply n times larger:

$$A^+_{(m,n)}=n\sum_{t=-\infty}^{\infty}\sum_{j=0}^{n-1}[F(m^+_{tn+j},m^+_{tn+j+1})-F(x_j,x_{j+1})]\ . \tag{8.23}$$

Similarly, the area under the lower separatrix is

$$A^-_{(m,n)}=-n\sum_{t=-\infty}^{\infty}\sum_{j=0}^{n-1}[F(m^-_{tn+j},m^-_{tn+j+1})-F(x_j,x_{j+1})]\ . \tag{8.24}$$

The final result is that the area in the (m,n) resonance is

$$A_{(m,n)}=A^+_{(m,n)}-A^-_{(m,n)}\ . \tag{8.25}$$

If we write (8.25) out explicitly, then it is similar to the general result (8.16) with $u=m^+$, $v=x$, $w=m^-$, and $z=x$. It may seem surprising that the contributions to the area of the resonance from the action of the (m,n) minimizing orbit add together instead of canceling; but this is so and comes from the fact that the asymptotic motion approaches the periodic orbit from the left in the upper separatrix and from the right in the lower separatrix.

In the above analysis, it has been assumed that there is only one minimizing (m,n) orbit. If there is more than one such orbit, then each gives a family of gaps, and one has to sum the contributions from each family.

3. Mean energy area formulas

The area formulas can also be obtained from the "mean energy," defined as a function of ω on the minimizing orbits as

$$L(\omega)=\lim_{t\to\infty}\frac{1}{2t}\sum_{j=-t}^{t-1}F(x_t,x_{t+1})|_{x_t\in M_\omega}\ . \tag{8.26}$$

Aubry (1982) shows that this is a convex function of ω, which implies that it has left and right derivatives and that they are equal almost everywhere. However, these derivatives differ at each rational value of ω. In fact, by considering limits of periodic orbits approaching homoclinic orbits or cantori, Chen (1987) has shown that these derivatives give the area functions

$$A^{\pm}_{(m,n)}=\left.\frac{d^{\pm}L}{d\omega}\right|_{\omega=m/n}$$
$$A_\omega=\left.\frac{dL}{d\omega}\right|_{\omega\ \text{irrational}}\ . \tag{8.27}$$

Here the $\pm$ in the derivative indicates that the derivative is taken from the right or left, respectively. These formulas are obtained by constructing the derivatives as limits of the difference $L(\omega')-L(\omega)$ as ω' approaches ω on minimizing periodic orbits, and by showing that the result is one of our previous area formulas. For irrational frequency, this formula gives the area under the cantorus partial barrier (providing the cantorus is hyperbolic), or, if one exists, under the invariant circle. We have no other formula for the area under the invariant circle in terms of the action of a finite number of orbits.

4. Resonances fill space

The twist condition implies that the minimizing orbits are ordered according to frequency along the vertical direction. Thus the area under the partial barriers as a function of frequency, A_ω, is a monotonically increasing function. One could think of it as the action for a nonintegrable system. Across every rational value, the area jumps by the amount (8.25), which is the area of the resonance. Thus A_ω is a *devil's staircase* [actually, the inverse function $\omega(A)$ is the devil's staircase].

Aubry (1982) has conjectured that this devil's staircase

is complete when there are no invariant circles and every cantorus is hyperbolic. Completeness means that the entire variation of the function is due to the jumps. Since the jumps each represent the area of a resonance, this implies that the resonances fill phase space. This can be numerically verified for the standard map when $k > k_{cr}(\gamma)$ (MacKay *et al.*, 1987) and analytically verified for the sawtooth map (Aubry, 1983a; Chen and Meiss, 1989).

Thus the resonances give a complete partition of irregular components. In Fig. 57 we show the (1,4), (1,3), and (2,5) resonances for the standard map. The partition of phase space resembles a tessellation—a tiling into n islands for each resonance, though the shapes of the tiles vary. There is a rough self-similarity apparent in the figure; for example, between the (1,3) and (2,5) resonances are exactly eight empty regions, which is just the right number for the (3,8) resonance, the Farey daughter of (1,3) and (2,5). Between the (3,8) resonance and each of its parents are just the right number of spaces for the next rationals on the Farey tree, (4,9) and (5,13). This structure continues for all levels. The structure is even more apparent for the sawtooth map (Chen *et al.*, 1990).

We shall use the resonance partition to construct a transport theory in the next section.

IX. TRANSPORT

In this section we develop simple models of transport based on flux between regions that partition phase space (recall the discussion of transport and flux in Sec. II.D). The ultimate goal, only partially attained to date, is to produce models that predict transport rates and provide explanations for such phenomena as the long-time tails seen in correlation functions.

We begin with an exact description of the transport process, based on the resonance partition.

A. Partitions

1. Resonances

Any annular irregular component is bounded by rotational invariant circles (Birkhoff's theorem, Sec. IV.C) and can be partitioned completely into rotational resonances (Sec. VIII.D). A resonance is the region of phase space enclosed by the separatrices and, up to the choice of a homoclinic point at which to switch from unstable to stable manifolds, is uniquely defined. We denote the resonance area by $A_{(m,n)}$. Each resonance has turnstiles in its upper and lower separatrices; their areas are denoted $\Delta W^{\pm}_{(m,n)}$ (see Fig. 58).

Since the rationals are countable, the resonances give a countable partition of phase space.

The transport process consists of the movement of trajectories among the resonances. Suppose that each resonance has an initial population of $N_{(m,n)}$ points. The goal of our transport description is to determine the popula-

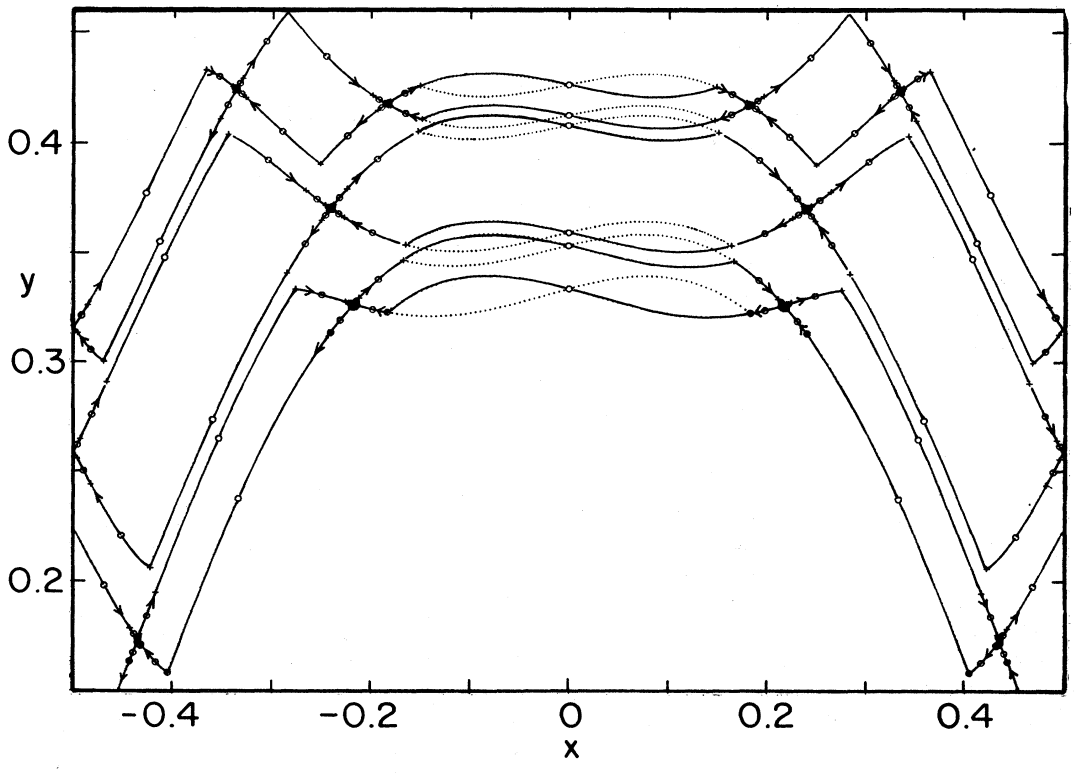

FIG. 57. Resonances for the standard map at $k = 1.283$.

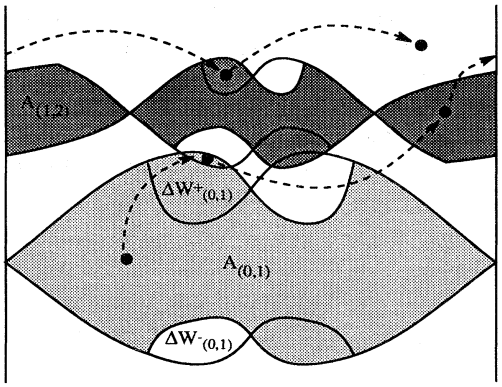

FIG. 58. Resonance partition of phase space. The upper turnstile of the (0,1) resonance overlaps the lower turnstile of the (1,2) resonance; so a direct transition is possible.

tion of each resonance after t iterations (MacKay *et al.*, 1987; Dana *et al.*, 1989).

In order to leave the (m,n) resonance, a point must fall in the exit lobe of either the upper or lower turnstile. Since there is a turnstile in only one island of the chain of n, points must move in a regular fashion through each island in the chain before finding themselves in the "principal" island with the turnstile (Fig. 58). Only when a point is in the turnstile in the principal island can a transition occur. Thus when a point enters a resonance, it must remain in the resonance for some multiple of n iterations.

A direct transition from a resonance (m,n) to (m',n') is possible only if the exit lobe of an (m,n) turnstile overlaps with the entry lobe of an (m',n') turnstile. In general, since the turnstiles have some finite height and the twist condition implies that resonances are ordered vertically according to frequency, the (m,n) turnstiles overlap with all resonances in some frequency range $\omega_L < m/n < \omega_U$ (Chen *et al.*, 1987). For example, in Fig. 58, since the upper turnstile of the (0,1) resonance partially overlaps the lower turnstile of the (1,2) resonance, it must overlap (completely) the turnstiles of all the resonances between 0/1 and 1/2.

2. Transport on a tree

The resonance partition and its corresponding transport description may be sufficient for some purposes. However, because the motion within a resonance is not typically featureless chaos, it may be important to consider partitioning the resonance itself for transport calculations. In fact, following the discussion of Sec. II.C, a rotational resonance can be partitioned into librational (class-1) resonances. Thus once a trajectory enters a class-zero resonance, it can get trapped in a sequence of class-1 resonances. Each of these has within it class-2 resonances. Thus the transport process occurs on a tree, whose branches correspond to the classes and whose leaves correspond to the stable islands, which are ultimately inaccessible (MacKay *et al.*, 1984; Meiss and Ott, 1986).

B. Markov models

1. Transition probabilities

The ultimate goal of a transport theory is to develop an approximate, statistical description of the motion (recall Sec. II.D). Recognizing the extreme complication of chaotic motion, we abandon the hope for an exact description of each trajectory and consider ensembles of trajectories. The simplest statistical model is a Markov model.

A Markov model consists of a partition of phase space into regions and a transition matrix, P_{ij}, which is the probability of a transition from region j to region i in one step. The entire motion is described in terms of the sequence of regions visited. When a trajectory is in one region, it has a given, fixed probability of making a transition to another region *independent* of its history. This is an essential assumption to the Markov model—that P_{ij} is independent of the past of the trajectory; it is almost certainly not true unless the partition is chosen with extreme care. We would like to investigate to what extent a partition into resonances has the Markov property.

Given that the population of the jth region at a given time is N_j, the Markov evolution states that upon one iteration the new distribution becomes

$$N_i' = \sum_j P_{ij} N_j \ . \tag{9.1}$$

It is known that a Markov partition exists for components on which the Lyapunov exponents are nonzero almost everywhere (Pesin, 1977). However, the construction of such a partition is nontrivial.

For the regions it is natural to choose resonances. Since the resonance partition is countable, the transition-probability matrix is discrete, but infinite in size. In the Markov approximation, the transition probability between two resonances is

$$P_{ij} = \frac{\mathcal{I}_{ij}}{A_j} \ , \tag{9.2}$$

where $\mathcal{I}_{ij}$ is the overlap area of the ith and jth resonance turnstiles. This transition probability is indeed exact for one iterate of the map; however, it is typically only qualitatively correct for longer times.

This kind of picture is in distinct contrast to the smoothed "diffusion" discussed in Sec. II.D. We would expect that the discrete model would be much more appropriate when there are partial barriers whose fluxes are small. This model should limit to the diffusive picture in the limit of large k.

A detailed construction of such a transport model can be given for the sawtooth map, which is almost every-

where hyperbolic (Dana *et al.*, 1989; Chen *et al.*, 1990). The Markov model works well for moderate values of k and does give the appropriate diffusive limit. In general we expect that if the system is chaotic enough—if the trapped set inside a resonance is a horseshoe—then the Markov model will work. However, in the typical case of a system with mixed regular and irregular components, the Markov model provides only a qualitative description.

2. Onset of transport near $k_{cr}(\gamma)$

The time for crossing an invariant circle is infinite. Above the critical value $k_{cr}(\omega)$ the flux through the cantorus grows smoothly from zero as the parameter is changed; thus we expect that the crossing time will have a singularity as the parameter limits to $k_{cr}(\omega)$ from above. Indeed, Chirikov observed in a numerical experiment for the transition from $y\sim 0$ to $y\sim\frac{1}{2}$ in the standard map (Chirikov, 1979b) that the transition time obeyed a power law

$$T\sim(k-k_{cr})^{-\eta}\ . \tag{9.3}$$

For the case of noble frequency, ΔW, and hence the flux, grows as a power law (MacKay, 1982)

$$\mathcal{F}\propto(k-k_{cr}(\omega))^{\eta},\quad \eta\approx 3.012\ . \tag{9.4}$$

Since the area of the connected chaotic component does not vanish in the neighborhood of k_{cr}, one would expect that the exponents in (9.3) and (9.4) should be identical. This has been numerically verified to a high degree for the standard map and the golden cantorus (Dana and Fishman, 1985) in the range $1<k<2.5$.

C. Escape from a resonance

In this section we consider the problem of escape from a single resonance with upper and lower turnstiles ΔW^{+} and ΔW^{-} and area A. Suppose that at $t=0$ the resonance is uniformly populated with N_0 particles, and the rest of phase space is empty. At the first step of the map a fraction of exactly

$$p=\frac{\Delta W^{+}+\Delta W^{-}}{A} \tag{9.5}$$

escapes from the resonance. Thus there are

$$N_1=(1-p)N_0 \tag{9.6}$$

particles remaining. Computing the fraction that escapes at the next iteration is more difficult, because the population in the resonance is no longer uniform. In the Markov approximation we assume that the N_1 particles have spread more or less uniformly throughout the resonance, so that a coarse graining would then view the resonance as uniformly populated with a lower density than at first. In this case, after t steps, we would have a population $N_t=(1-p)^t N_0$ remaining in the resonance, and at the tth step,

$$N_{t-1}-N_t=p(1-p)^{t-1}N_0 \tag{9.7}$$

particles escape. Thus the population would decay exponentially with time with escape rate

$$r=-\log(1-p)\ .$$

Exponential decay is indeed often observed, especially when the interior of the resonance has no apparent elliptic regions.

1. Transit-time decomposition

To obtain an exact description it is necessary to follow the iterates of the incoming lobes of the turnstiles [i.e., *lobe dynamics* (Rom-Kedar and Wiggins, 1988)]. Let $\mathcal{I}$ represent the collection of incoming lobes, $\mathcal{E}$ the exiting lobes, and $\mathcal{I}_t$ the tth iterate of $\mathcal{I}$. The areas of these regions are denoted $\mu(\mathcal{I}_t)$, etc.

Since the area of the exiting and of the entering lobes is the area of the turnstiles, we have

$$\mu(\mathcal{I})=\mu(\mathcal{I}_t)=\mu(\mathcal{E})=\Delta W^{+}+\Delta W^{-}\ . \tag{9.8}$$

The set that enters the resonance at $t=0$ is in the region $\mathcal{I}_1$ at $t=1$ (see Fig. 59). The fraction of $\mathcal{I}_1$ that intersects $\mathcal{E}$ exits the resonance upon the next iterate; and so we say it has a *transit time* of 1. In general, the set that traverses the resonance is exactly t steps is

$$\mathcal{T}_t=\mathcal{E}\cap\mathcal{I}_t\ . \tag{9.9}$$

The sets $\mathcal{T}_t$ are clearly disjoint. Furthermore,

$$\mu(\mathcal{E})=\sum_{j=0}^{\infty}\mu(\mathcal{T}_j)\ , \tag{9.10}$$

because any area that enters must eventually leave (recall Sec. II.B). The *transit-time decomposition* of $\mathcal{E}$ is the decomposition into the sets $\mathcal{T}_t$.

The transit-time decomposition of the exit set can be analyzed completely in terms of various homoclinic orbits (Easton, 1991). Figure 60 sketches the decomposi-

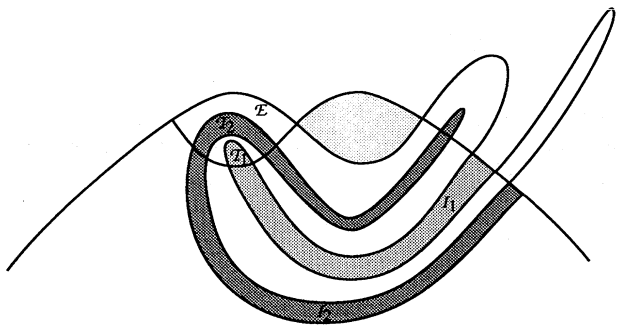

FIG. 59. Transit time for a resonance. Only the upper turnstile is shown.

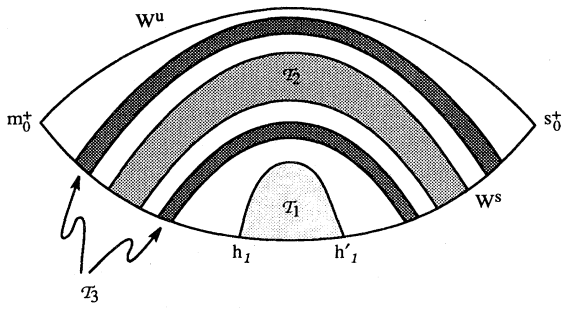

FIG. 60. Transit-time decomposition of an exit set.

tion of the exit lobe of the upper turnstile. It is bounded by points on the upper minimizing and minimax homoclinic orbits, m_0^+ and s_0^+. The area of the region $\mathcal{T}_1$ is the area contained between the segments of stable and unstable manifolds of the two homoclinic orbits h_1 and h_1', shown in the figure. We recall that area contained between the stable and unstable manifolds of any homoclinic pair is given by Eq. (8.15). For $\mathcal{T}_2$ there are four homoclinic orbits that must be computed, and the area is given by Eq. (8.16). In general, each $\mathcal{T}_j$ consists of a set of strips stretching across $\mathcal{E}$, and possibly of some lobes that do not traverse $\mathcal{E}$ entirely. However, all these regions can be computed by knowing the actions of various homoclinic points on the segment of stable manifold between m_0 and s_0.

2. Lobe dynamics

The transit-time problem is closely related to the exit time problem, since the population escaping at time t is the occupied portion of the outgoing turnstile. To calculate this, one must subtract the total area of the transiting regions,

$$\text{Area escaping at time } t = \mu(\mathcal{E}) - \sum_{j=0}^{t-1} \mu(\mathcal{T}_j) \ . \tag{9.11}$$

This implies that the number of particles escaping at time t is

$$N_{t-1} - N_t = \left[\mu(\mathcal{E}) - \sum_{j=0}^{t-1} \mu(T_j) \right] \frac{N_0}{A} \ , \tag{9.12}$$

since the density in the occupied region, N_0/A, is invariant under the map. Using Eqs. (9.5), (9.8), and (9.10), we can write this as

$$N_{t-1} - N_t = pN_0 \sum_{j=t}^{\infty} \frac{\mu(\mathcal{T}_j)}{\mu(\mathcal{E})} \ . \tag{9.13}$$

Formulas similar to (9.13) and also applicable to more general cases can be given (Rom-Kedar and Wiggins, 1988; Rom-Kedar, 1990; Beigie *et al.*, 1991). Unfortunately, these formulas give no indication as to the size of the $\mu(\mathcal{T}_t)$, nor how to compute them.

If the escape were a pure exponential, then the area of the transit sets would have to decrease exponentially with time:

$$\mu(\mathcal{T}_t) \approx \mu(T_1)\alpha^{t-1} \ ; \tag{9.14}$$

and using (9.10) in (9.13) would yield

$$N_t - N_{t-1} = p\alpha^{t-1} N_0 \ . \tag{9.15}$$

Comparing this with (9.7), we see that the Markov escape rate would be exact if $\alpha = 1-p$, or

$$\frac{\mu(\mathcal{T}_1)}{\mu(\mathcal{E})} = \frac{\mu(\mathcal{E})}{A} \ , \tag{9.16}$$

which is a kind of "mixing" assumption: the fraction of area that transits the resonance is the same as what we would expect if the incoming turnstile were completely mixed throughout the resonance area.

When the parameter k of the standard map is large enough, we expect—and observe numerically—exponential decay. Even so, the rate α is often different from $1-p$. Furthermore, it often happens that it takes several iterations before the incoming turnstile intersects $\mathcal{E}$, so that $\mu(T_t)=0$ for $t < t_m$. In this case the number of particles decreases linearly for $t < t_m$, and then exponentially thereafter. In this case one need only compute t_m and $\mu(\mathcal{T}_{t_m})$ in order to calculate α.

For the sawtooth map, which can be analyzed completely because it is piecewise linear (Percival and Vivaldi, 1987a, 1987b), the rates can be computed analytically in some cases. For example, when $k > \frac{4}{3}$, (9.16) is valid for the (0,1) resonance and the Markov model is exact (Chen *et al.*, 1990). More generally, whenever the Lyapunov multiplier of the minimizing (m,n) orbit is larger than 3^n, the Markov model is exact.

3. Periodic orbit theory

An alternative theory for escape rates for systems that are hyperbolic can be obtained by analyzing orbits that are trapped in the resonance forever (Kadanoff and Tang, 1984; Grebogi *et al.*, 1988). Consider a small box about a point on a hyperbolic period t trapped orbit. After t iterations the box returns to the neighborhood of the initial point; but due to stretching along the unstable direction by a factor λ, the eigenvalue of the Jacobian matrix (2.15), only a fraction of $1/\lambda$ of the iterate overlaps with the original box. The escape from the neighborhood of the orbit is exponential.

In order for an orbit to remain trapped in the resonance for a long time, it must be close to the trapped orbits; after t steps the trapped set is a small neighborhood of the trapped period t orbits. The fraction remaining in the resonance after t steps is proportional to the sum of the fractions remaining near each of these orbits:

$$N_t \sim \sum_j^{(t)} \frac{1}{\lambda_j} . \tag{9.17}$$

Here λ_j is the eigenvalue of the jth trapped period t orbit, and the sum is over all trapped period t orbits. In the hyperbolic case the escape should be exponential; so as $t \to \infty$, (9.17) is proportional to e^{-rt}, with the escape rate r.

For the simplest case, all trapped orbits have the same Lyapunov multiplier λ; so $\lambda_j = \lambda^t$. Then the escape rate is

$$r = \log(\lambda) - \mathrm{ent}(T) ,$$

where $\mathrm{ent}(T)$ is the metric entropy of the map T, that is, the growth rate of the number of periodic orbits of period t. This formula works well for the sawtooth map, even when $k < \frac{4}{3}$ (Chen *et al.*, 1990).

More generally, the sum (9.17) must be evaluated by computing the periodic orbits. Considerable savings in computational effort can be obtained by reordering the sum to take advantage of the fact that long period orbits can be closely approximated by products of shorter orbits (Artuso *et al.*, 1990a, 1990b). Often escape rates can be obtained with high accuracy using only short orbits.

A similar analysis yields formulas for the diffusion coefficient in terms of periodic orbits (Dana, 1989; Cvitanovic and Eckmann, 1991).

Unfortunately the periodic orbit formulation seems unable to deal with systems that are not hyperbolic. In such cases the decay is not exponential—in fact, observations imply it is algebraic.

4. Algebraic decay

Whenever there are elliptic islands within a region, the escape rates are not exponential, but rather appear to be algebraic (Chirikov, 1983; Karney, 1983; Chirikov and Shepelyanksy, 1984; Geisel *et al.*, 1987; Petschel and Geisel, 1991).

One of the major outstanding questions in this field is how to explain this behavior from first principles.

Numerical experiments (Karney, 1983) show that the longest orbits are those that get trapped arbitrarily close to the outermost invariant circles surrounding elliptic islands, the *boundary circles*. From our viewpoint, this is not unexpected, since the flux through orbits limiting on an invariant circle approaches zero [recall (8.18)]; however, this effect must compete with the concomitant decrease in the area of the regions.

In fact, the area of the turnstiles decreases more rapidly than that of the resonances. This is because there is only one turnstile in a chain of n islands. Moving closer to the boundary circle corresponds to the frequency becoming a better approximation to the frequency of the irrational boundary; thus the closer island chains have longer periods. Furthermore, as one moves closer to the boundary, there is an approximate geometric scaling of the structures: an island of the closer chain is similar to an island of the farther one. Thus the turnstile area scales as the area of one island. The area of the entire resonance, however, scales as the period times the area of one island, and thus decreases more slowly.

These qualitative statements can be made precise for the case of boundary circles (Greene *et al.*, 1986). Suppose

$$\frac{m_i}{n_i} \to \omega$$

represents the Farey sequence of rationals on the chaotic side of the boundary circle. For example, if the chaotic component is an annulus below the boundary circle, then $m_i/n_i < \omega$. One observes that the turnstiles in the (m_i, n_i) resonances scale as

$$\Delta W_{i,i+1} \approx \frac{\Delta W_0}{n_i^\beta}, \quad \beta = 3.05 . \tag{9.18}$$

This reduces to (8.18) for the case of noble circles where $n_i \sim \gamma^{2i}$. The area of a single island scales in the same way. However, the area of the resonance is n_i times the area of one island; thus

$$A_i \approx \frac{A_0}{n_i^{\beta-1}} . \tag{9.19}$$

Therefore the transition probabilities (9.2) have the scaling

$$p_{i+1,i} \approx \frac{p_0}{n_i} , \tag{9.20}$$

and the ratio of the probability for a transition towards the boundary circle to one away from the boundary circle is

$$\frac{p_{i+1,i}}{p_{i-1,i}} = \frac{\Delta W_{i,i+1}}{\Delta W_{i-1,i}} \approx \left[\frac{n_i}{n_{i+1}} \right]^\beta . \tag{9.21}$$

If we assume that n_i grows geometrically, as it does for noble numbers, then the transition probabilities have a geometrical scaling. Using these in the model (9.1) results in an infinite, self-similar Markov chain with nearest-neighbor connections. This model can be solved exactly (Hanson *et al.*, 1985). It predicts an algebraic decay of the number of trapped particles at the rate

$$N_t \sim t^{-z}, \quad z = 1 + \beta , \tag{9.22}$$

where β is given in (9.18). Numerically observed decays are much slower.

The model can be improved by including the branches of the tree and developing a scaling for transitions from one branch of the tree to another (Meiss and Ott, 1986). Using this analysis and plausible values for the scaling coefficients, one finds that the decay exponent z becomes 2.96—which still does not agree with numerical rates.

One possible resolution of this discrepancy is that the use of the universal scaling relations (9.18) and (9.19) is

inappropriate in a comparison with numerical experiments (Murray, 1991). In particular, these are valid only in a tiny neighborhood of the critical circle—a neighborhood which would be reached by a typical orbit only after many iterations; Murray estimates 10^{10} for one case. Murray argues that farther from the critical circle the scalings become $\Delta W \propto n_i^{-2\beta}$, and $A_i \propto n_i^{-2}$. This would give an exponent in (9.22) of $z \sim 1.45$, which agrees more closely with the experiments.

Given the assumptions that are required for the Markov model, these results must be taken as incomplete. The numerical experiments are also not totally convincing. Experiments require iterating the map at least 10^6 steps to detect algebraic decay; the record number of iterates is 10^{12} (Karney, 1983). Great care must be used in interpreting these results, since the Lyapunov multipliers for the orbits involved imply that all accuracy in the computation is lost. Karney used an integer representation to ensure that his map was computationally one to one; however, recent number theoretic results imply that great care must be used in such discretizations, in order that the discrete map be a good representation of the continuous map (Percival and Vivaldi, 1987a).

ACKNOWLEDGMENTS

This paper arose from a series of lectures I first gave at the dynamics seminar at UC Berkeley during the spring semester of 1989. I would like to thank Allan Kaufman for inviting me to visit. It was at the instigation of Allan Kaufman, Edgar Knobloch, and Robert Littlejohn that I gave these lectures, and I appreciate their probing questions as well as those from Allan Lichtenberg, Mike Lieberman, and Dan Goroff. They have certainly improved the clarity of my presentation. My understanding of this subject has grown out of an extremely fruitful collaboration with Ian Percival and Robert MacKay, and I used the lecture notes of MacKay and Stark (1985) heavily in preparing my lectures. I would also like to thank Robert Easton for helpful comments on the manuscript and for many useful conversations.

This work was supported by the Department of Energy under Contract No. DE-AC03-76SF00098 and by the National Science Foundation under Grant No. DMS-9001103.

APPENDIX A: DIFFERENTIAL FORMS

A differential n form is an object that operates on n vectors to give a real number (Arnol'd, 1978). A one-form, α, is analogous to a covariant vector, it acts on an ordinary vector $\mathbf{v}$ with the dot product to give a real: $\alpha(\mathbf{v})=\alpha_i v^i$ (we use the summation convention). For example, the form df is a covariant vector associated with the gradient of a function f; operating on a vector $\mathbf{v}$ with df gives the derivative of f in the direction of $\mathbf{v}$:

$$df(\mathbf{v})=v^i\frac{\partial f}{\partial x^i}\ . \tag{A1}$$

Associated with the coordinate function x^i is a one-form dx^i. The action of dx^i on a vector $\mathbf{v}$ is v^i, the ith component of $\mathbf{v}$.

A two-form, ω, is an equivalent to an antisymmetric matrix, say, ω_{ij}—its action on two vectors is $\omega(\mathbf{u},\mathbf{v})=u^i\omega_{ij}v^j$. Antisymmetry implies that $\omega(\mathbf{u},\mathbf{v})=-\omega(\mathbf{v},\mathbf{u})$. The form with which we are most concerned is the symplectic form $\omega=\sum dp^i\wedge dq^i$. The result of acting on two vectors with ω is the number

$$\omega(\mathbf{u},\mathbf{v})=\sum_i dp^i(\mathbf{u})\,dq^i(\mathbf{v})-dp^i(\mathbf{v})\,dq^i(\mathbf{u})\ , \tag{A2}$$

which can be interpreted as an area [recall Eq. (1.15) and Fig. 3].

The exterior derivative d converts an n form to an $n+1$ form. Thus the exterior derivative of pdq is the two-form ω. If the exterior derivative of a form vanishes, the form is said to be *closed*. If an n form can be written as the exterior derivative of an $n-1$ form, it is said to be *exact*. The exterior derivative of an exact form is zero; so it is closed.

Differential n forms can be integrated over n-dimensional surfaces. For example, choose an arbitrary two-dimensional surface $\mathcal{S}$ embedded in a $2N$-dimensional manifold. Associate an orientation to $\mathcal{S}$ by choosing a direction to traverse the boundary of $\mathcal{S}$. The integral

$$A=\sum_{i=1}^{N}\int_{\mathcal{S}} dp^i\wedge dq^i \tag{A3}$$

is a sum over the projected areas of the surface $\mathcal{S}$ onto the canonical planes; the wedge product means that the areas are positive if the projection of the boundary is traversed clockwise, or negative if counterclockwise. The generalization of Stokes's theorem to n dimensions implies that since $\omega=d(pdq)$, the integral (A3) can be written as the integral of pdq over the boundary of $\mathcal{S}$

$$A=\sum_{i=1}^{N}\int_{\partial\mathcal{S}} p^i dq^i\ . \tag{A4}$$

APPENDIX B: CIRCLE MAPS

Here we review a few basic facts about homeomorphisms of the circle (Cornfeld *et al.*, 1982, pp. 73–95). Let $\alpha(x)$ be a continuous, monotonic increasing function of x satisfying $\alpha(x+1)=\alpha(x)+1$ (see, for example, Fig. 43).

B1 Lemma. *There exists an ω such that for all $x\in\mathbb{R}$ and integers (m,n)*

$$\begin{aligned} n\omega > m &\Longrightarrow \alpha^n(x)-m>x\ , \\ n\omega < m &\Longrightarrow \alpha^n(x)-m<x\ . \end{aligned} \tag{B1}$$

The ω that satisfies this lemma is the rotation number of α. An important consequence of this lemma is that the orbit cannot deviate too far from uniform rotation. To show this from the above two inequalities, in the first case let m be the greatest integer less than $n\omega$, and in the second let m be the smallest integer greater than $n\omega$; then we can bound the difference

$$|\alpha^n(x)-x-n\omega| \leq 1 \ . \tag{B2}$$

Equation (B2) implies

B2 Theorem. *The limit*

$$\omega = \lim_{n\to\infty} \frac{\alpha^n(x)}{n} \tag{B3}$$

exists and does not depend on the choice of $x \in \mathbb{R}$. *The rotation number* ω *is rational only if some power of* α *has a fixed point.*

Choose an arbitrary point x_0 and consider its trajectory under α. Let Ω be the set of *limit points* of the orbit: $x \in \Omega$ if there is a sequence $x_j = \alpha^j(x_0)$ such that $x_j \to x$ as $j \to \infty$. By definition Ω is closed. Then

B3 Theorem (Poincaré, 1885; Denjoy, 1932). *If* ω *is irrational,*

(a) Ω *is independent of the choice of* x_0;

(b) Ω *is invariant;*

(c) Ω *is either the entire circle or is a Cantor set.*

A *Cantor set* C is a nonempty, perfect, totally disconnected, compact set:

Perfect $\Longleftrightarrow$ Every point in the set is a limit point of other points in the set: For all $x \in C$, there is a sequence $x^{(n)} \in C$ such that $x^{(n)} \neq x$ and $x^{(n)} \to x$ as $n \to \infty$.

Totally disconnected $\Longleftrightarrow$ For any $x, y \in C$, such that $x \neq y$, C can be written as the union of disjoint, closed sets A and B for which $x \in A$ and $y \in B$.

Compact $\Longleftrightarrow$ Every sequence $x^{(j)}$ in C has a convergent subsequence.

This definition of the Cantor set is a purely topological one. It does not require the set to be embedded in any other space. The standard example of a Cantor set is a subset of the interval [0,1], Fig. 61. Remove the open interval $(\frac{1}{3}, \frac{2}{3})$, leaving two closed intervals. Remove the middle third from each of these. Continue this procedure *ad infinitum*. This construction shows that the complement of a Cantor set is a countable set of gaps.

A Cantor set contains an uncountable number of points; in particular, there are points that are not the end point of any gap. To see this for the middle-thirds example, we code the points in the set in a base-three representation. The geometrical construction of the middle-thirds set implies that it consists of points whose base-three representations have no 1's, i.e., 0.2022202000. . . . There are an uncountable number of such sequences. By contrast, points that lie on the end points of a gap have finite base-three expansions, since they are rationals with powers of 3 in the denominator.

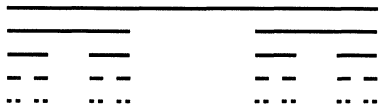

FIG. 61. Three levels in the construction of the middle-thirds Cantor set.

The Hausdorff dimension of a Cantor set embedded in some manifold can take any value. In the middle-thirds case, the Hausdorff dimension is log(2)/log(3). If the fraction removed at each level is increased then this dimension decreases. The invariant Cantor sets arising in the twist map case (cantori) typically have zero Hausdorff dimension.

REFERENCES

Abramowitz, M., and I. A. Stegun, 1965, *Handbook of Mathematical Functions* (Dover, New York).

Aref, H., 1984, "Stirring by Chaotic Advection," J. Fluid Mech. **143**, 1–21.

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

Arnol'd, V. I., and A. Avez, 1968, *Ergodic Problems of Classical Mechanics* (Benjamin, New York).

Arrowsmith, D. K., and C. M. Place, 1990, *An Introduction to Dynamical Systems* (Cambridge University, Cambridge).

Artuso, R., E. Aurell, and P. Cvitanovic, 1990a, "Recycling of Strange Sets I: Cycle Expansions," Nonlinearity **3**, 325–360.

Artuso, R., E. Aurell, and P. Cvitanovic, 1990b, "Recycling of Strange Sets II: Applications," Nonlinearity **3**, 361–386.

Aubry, S., 1978, "The new concept of transitions by breaking of analyticity in a crystallographic mode" in *Solitons and Condensed Matter Physics*, Springer Series in Solid-State Sciences Vol. 8, edited by A. R. Bishop and T. Schneider (Springer-Verlag, New York), pp. 264–277.

Aubry, S., 1982, "The Devil's Staircase Transformation in Incommensurate Lattices," in *The Rieman Problem, Complete Integrability and Applications* Lecture Notes in Mathematics 925, edited by D. Chudnovsky and G. Chudnovsky (Springer-Verlag, New York), pp. 221–245.

Aubry, S., 1983a, "Exact Models with a Complete Devil's Staircase," J. Phys. C **16**, 2497–2508.

Aubry, S., 1983b, "The Twist Map, the Extended Frenkel-Kontorova Model and the Devil's Staircase," Physica D **7**, 240–258.

Aubry, S., and P. Y. Le Daeron, 1983, "The Discrete Frenkel-Kontorova Model and Its Extensions," Physica D **8**, 381–422.

Bangert, V., 1988, "Mather Sets for Twist Maps and Geodesics on Tori," Dyn. Rep. **1**, 1–56.

Beigie, D., A. Leonard, and S. Wiggins, 1991, "A Global Study of Enhanced Stretching and Diffusion in Chaotic Tangles,"

Phys. Fluids A **3**, 1039–1050.
Bensimon, D., and L. P. Kadanoff, 1984, "Extended Chaos and Disappearance of KAM Trajectories," Physica D **13**, 82–89.
Berry, M. V., 1982, "Regularity and Chaos in Classical Mechanics, Illustrated by Three Deformations of a Circular Billiard," Eur. J. Phys. **2**, 91–102.
Birkhoff, G. D., 1913, "Proof of Poincaré's Geometric Theorem," Trans. Am. Math. Soc. **14**, 14–22.
Birkhoff, G. D., 1920, "Surface Transformations and Their Dynamical Applications," Acta Math. **43**, 1–119.
Birkhoff, G. D., 1935, "Nouvelles Recherches sur les Systemes Dynamiques," Memoriae Point. Acad. Sci. Novi Lyncaei **1**, 85–216.
Boozer, A. H., and R. B. White, 1982, "Particle Diffusion in Tokamaks with Partially Destroyed Magnetic Surfaces," Phys. Rev. Lett. **49**, 786–789.
Bruschi, M., O. Ragnisco, R. M. Santini, and T. Gui-Zhang, 1991, "Integrable Symplectic Maps," Physica D **49**, 273–294.
Bunimovich, L. A., 1974, "On the Ergodic Properties of Certain Billiards," Funct. Anal. Appl. **8**, 254–255.
Carrigan, R. A., F. R. Huson, and M. Month, 1982, Eds., *Physics of High Energy Particle Accelerators,* AIP Conference Proceedings No. 87 (AIP, New York).
Cary, J. R., 1984, "Construction of Three-Dimensional Vacuum Magnetic Fields with Dense Nested Flux Surfaces," Phys. Fluids **27**, 119–128.
Cary, J. R., and R. G. Littlejohn, 1983, "Noncanonical Hamiltonian Mechanics and Its Application to Magnetic Field Line Flow," Ann. Phys. (NY) **151**, 1–34.
Cary, J. R., J. D. Meiss, and A. Bhattacharjee, 1981, "Statistical Characterization of Periodic, Area-Preserving Mappings," Phys. Rev. A **23**, 2744–2746.
Cassels, J. W. S., 1965, *An Introduction to Diophantine Approximation* (Cambridge University, Cambridge).
Chen, Q., 1987, "Area as a Devil's Staircase in Twist Maps," Phys. Lett. A **123**, 444–450.
Chen, Q., I. Dana, J. D. Meiss, and I. Percival, 1990, "Resonances and Transport in the Sawtooth Map," Physica D **46**, 217–240.
Chen, Q., and J. D. Meiss, 1989, "Flux, Resonances and the Devil's Staircase for the Sawtooth Map," Nonlinearity **2**, 347–356.
Chen, Q., J. D. Meiss, and I. C. Percival, 1987, "Orbit Extension Method for Finding Unstable Orbits," Physica D **29**, 143–154.
Chirikov, B. V., 1979a, "Homogeneous Model for Resonant Particle Diffusion in an Open Magnetic Confinement System," Sov. J. Plasma Phys. **5**, 492–497.
Chirikov, B. V., 1979b, "A Universal Instability of Many-Dimensional Oscillator Systems," Phys. Rep. **52**, 265–379.
Chirikov, B. V., 1983, "Chaotic Dynamics in Hamiltonian Systems with Divided Phase Space," in *Dynamical Systems and Chaos,* Lecture Notes in Physics Vol. 179, edited by L. Garrido (Springer-Verlag, Berlin), pp. 29–46.
Chirikov, B. V., and D. L. Shepelyanksy, 1984, "Correlation Properties of Dynamical Chaos in Hamiltonian Systems," Physica D **13**, 395–400.
Cornfeld, I. P., S. V. Fomin, and Y. G. Sinai, 1982, *Ergodic Theory,* Grundlehren der mathematischen Wissenschaften (Springer-Verlag, New York).
Cvitanovic, P., and J. P. Eckmann, 1991, "Transport Properties of the Lorentz Gas in Terms of Periodic Orbits," NORDITA preprint.
Dana, I., 1989, "Hamiltonian Transport on Unstable Orbits," Physica D **39**, 205–230.
Dana, I., and S. Fishman, 1985, "Diffusion in the Standard Map," Physica D **17**, 63–74.
Dana, I., N. Murray, and I. C. Percival, 1989, "Resonances and Diffusion in Periodic Hamiltonian Maps," Phys. Rev. Lett. **62**, 233–236.
Davis, M. J., 1985, "Bottlenecks to Intramolecular Energy Transfer and the Calculation of Relaxation Rates," J. Chem. Phys. **83**, 1016–1035.
de la Llave, R., and D. Rana, 1990, "Accurate Strategies for Small Divisor Problems," Bull. Am. Math. Soc. **22**, 85–90.
Denjoy, A., 1932, "Sur les Courbes Définés par les Équations Différentielles à la Surface du Tore," J. Math. Pures Appl. **11**, 333–375.
Devaney, R., 1976, "Reversible Diffeomorphisms and Flows," Trans. Am. Math. Soc. **218**, 89–113.
Devaney, R., 1986, *An Introduction to Chaotic Dynamical Systems* (Benjamin/Cummings, Menlo Park).
DeVogelaere, R., 1958, "On the Structure of Symmetric Periodic Solutions of Conservative Systems, with Applications," in *Contributions to the Theory of Nonlinear Oscillations,* edited by S. Lefshetz (Princeton University, Princeton, NJ), pp. 53–84.
Dragt, A. J., and J. M. Finn, 1976, "Insolubility of Trapped Particle Motion in a Magnetic Dipole Field," J. Geophys. Res. **81**, 2327–2340.
Easton, R. W., 1991, "Transport Through Chaos," Nonlinearity **4**, 583–590.
Evans, L. R., 1983, "The Beam-Beam Interaction," CERN Report No. SPS/83-38 (DI-MST).
Geisel, T., and S. Thomae, 1984, "Anomalous Diffusion in Intermittent Chaotic Systems," Phys. Rev. Lett. **52**, 1936–1939.
Geisel, T., A. Zacherl, and G. Radons, 1987, "Generic $1/f$ Noise in Chaotic Hamiltonian Dynamics," Phys. Rev. Lett. **59**, 2503–2506.
Gelfand, I. M., and S. V. Fomin, 1963, *Calculus of Variations* (Prentice-Hall, Englewood Cliffs, NJ).
Golé, C., 1991, "Monotone maps of $T^n \times R^n$ and their periodic orbits," in *The Geometry of Hamiltonian Systems,* edited by T. Ratiu (Springer-Verlag, New York), pp. 341–366.
Goroff, D. L., 1985, "Hyperbolic Sets for Twist Maps," Ergodic Theory Dyn. Syst. **5**, 337–339.
Grebogi, C., E. Ott, and J. A. Yorke, 1988, "Unstable Periodic Orbits and the Dimension of Multifractal Chaotic Attractors," Phys. Rev. A **37**, 1711–1724.
Greene, J. M., 1979, "A Method for Computing the Stochastic Transition," J. Math. Phys. **20**, 1183–1201.
Greene, J. M., H. Johannesson, B. Schaub, and H. Suhl, 1987, "Scaling Anomaly at the Critical Transition of an Incommensurate Structure," Phys. Rev. A **36**, 5858–5861.
Greene, J. M., R. S. MacKay, and J. Stark, 1986, "Boundary Circles for Area-Preserving Maps," Physica D **21**, 267–295.
Greene, J. M., R. S. MacKay, F. Vivaldi, and M. J. Feigenbaum, 1981, "Universal Behaviour in Families of Area-Preserving Maps," Physica D **3**, 468–486.
Hanson, J. D., J. R. Cary, and J. D. Meiss, 1985, "Algebraic Decay in Self-Similar Markov Chains," J. Stat. Phys. **39**, 327–345.
Hardy, G. H., and E. M. Wright, 1979, *An Introduction to the Theory of Numbers* (Oxford University, Oxford).
Hedlund, G. A., 1932, "Geodesics on a Two-Dimensional Riemannian Manifold with Periodic Coefficients," Ann. Math. **33**, 719–739.
Hénon, M., and C. Heiles, 1964, "The Applicability of the Third Integral of Motion: Some Numerical Experiments,"

Astron. J. **69**, 73–79.
Herman, M. R., 1983, "Sur les Courbes Invariantes par les Difféomorphismes de L'anneau. Vol. 1," Astérisque **103–104**, 1–221.
Herman, M. R., 1985, "Sur les Courbes Invariantes par les Difféomorphismes de L'anneau. Vol. 2," Astérisque **144**, 1–248.
Herman, M. R., 1988, "Existence et Non-existence de Tores Invariants par des Difféomorphismes Symplectiques," Ecole Polytechnique, Exposé XIV.
Ichikawa, Y. H., T. Kamimura, and T. Hatori, 1987, "Stochastic Diffusion in the Standard Map," Physica D **29**, 247.
Jowett, J. M., M. Month, and S. Turner, 1986, Eds., *Nonlinear Dynamics Aspects of Particle Accelerators,* Lecture Notes in Physics Vol. 247 (Springer-Verlag, Berlin).
Kadanoff, L. P., and C. Tang, 1984, "Escape from Strange Repellers," Proc. Natl. Acad. Sci. USA **81**, 1276–1279.
Karney, C. F. F., 1983, "Long Time Correlations in the Stochastic Regime," Physica D **8**, 360–380.
Karney, C. F. F., A. B. Rechester, and R. B. White, 1982, "Effect of Noise on the Standard Mapping," Physica D **4**, 425–438.
Katok, A., 1982, "Some Remarks on the Birkhoff and Mather Twist Map Theorems," Ergodic Theory Dyn. Syst. **2**, 185–194.
Ketoja, J. A., and R. S. MacKay, 1989, "Fractal Boundary for the Existence of Invariant Circles for Area-Preserving Maps: Observation and Renormalization Explanation," Physica B **35**, 318–334.
Khakhar, D. V., H. Rising, and J. M. Ottino, 1986, "Analysis of Chaotic Mixing in two Model Systems," J. Fluid Mech. **172**, 419–451.
Khinchin, A. Y., 1964, *Continued Fractions* (University of Chicago, Chicago).
Kook, H. T., and J. D. Meiss, 1989, "Periodic Orbits for Reversible, Symplectic Mappings," Physica D **35**, 65–86.
Landford, O. E., 1973, "Introduction to the Mathematical Theory of Dynamical Systems," in *Chaotic Behavior of Deterministic Systems,* edited by G. Ioos, R. H. G. Helleman, and R. Stora (North-Holland, Amsterdam), pp. 3–51.
Li, W., and P. Bak, 1986, "Fractal Dimension of Cantori," Phys. Rev. Lett. **57**, 655–658.
Lichtenberg, A. J., and M. A. Lieberman, 1982, *Regular and Stochastic Motion* (Springer-Verlag, New York).
MacKay, R. S., 1982, "Renormalization in Area-Preserving Maps," Ph.D. thesis (Princeton University).
MacKay, R. S., 1983, "A Renormalization Approach to Invariant Circles in Area-Preserving Maps," Physica D **7**, 283–300.
MacKay, R. S., 1986, "Transition to Chaos for Area-Preserving Maps," in *Nonlinear Dynamics Aspects of Particle Accelerators,* Lecture Notes in Physics Vol. 247, edited by J. M. Jowett, M. Month, and S. Turner (Springer-Verlag, Berlin), pp. 390–454.
MacKay, R. S., 1987, "Hyperbolic Cantori Have Dimension Zero," J. Phys. A **20**, No. 9, L559–L561.
MacKay, R. S., 1991, "On Greene's Residue Criterion," University of Warwick preprint.
MacKay, R. S., and J. D. Meiss, 1983, "Linear Stability of Periodic Orbits in Lagrangian Systems," Phys. Lett. A **98**, 92–94.
MacKay, R. S., and J. D. Meiss, 1987, *Hamiltonian Dynamical Systems: a reprint selection* (Adam-Hilger, London).
MacKay, R. S., J. D. Meiss, and I. C. Percival, 1984, "Transport in Hamiltonian Systems," Physica D **13**, 55–81.
MacKay, R. S., J. D. Meiss, and I. C. Percival, 1987, "Resonances in Area Preserving Maps," Physica D **27**, 1–20.
MacKay, R. S., J. D. Meiss, and J. Stark, 1989, "Converse KAM Theory for Symplectic Twist Maps," Nonlinearity **2**: 555–570.
MacKay, R. S., and I. C. Percival, 1985, "Converse KAM: Theory and Practice," Commun. Math. Phys. **98**, 469–512.
MacKay, R. S., and J. Stark, 1985, "Lectures on Orbits of Minimal Action for Area-Preserving Maps," Mathematics Institute, University of Warwick preprint.
Mather, J. N., 1982, "Existence of Quasi-Periodic Orbits for Twist Homeomorphisms of the Annulus," Topology **21**, 457–467.
Mather, J. N., 1984, "Non-Existence of Invariant Circles," Ergodic Theory Dyn. Syst. **2**, 301–309.
Mather, J. N., 1985, "More Denjoy Minimal Sets for Area Preserving Diffeomorphisms," Comment. Math. Helv. **60**, 508–577.
Mather, J. N., 1986, "A Criterion for Non-Existence of Invariant Circles," Publ. Math. I.H.E.S. **63**, 153–204.
McMillan, E. M., 1971, "A Problem in the Stability of Periodic Systems," in *Topics in Modern Physics, a Tribute to E. V. Condon,* edited by E. Brittin and H. Odabasi (Colorado University, Boulder), pp. 219–244.
Meiss, J. D., 1986, "Class Renormalization: Islands around Islands," Phys. Rev. A **34**, 2375–2383.
Meiss, J. D., J. R. Cary, C. Grebogi, J. D. Crawford, A. N. Kaufman, and H. D. I. Abarbanel, 1983, "Correlations of Periodic, Area-Preserving Maps," Physica D **6**, 375–384.
Meiss, J. D., and E. Ott, 1986, "Markov Tree Model of Transport in Area-Preserving Maps," Physica D **20**, 387–402.
Milnor, J., 1963, *Morse Theory,* Annals of Mathematical Studies Vol. 51 (Princeton University, Princeton, NJ).
Morse, M., 1924, "A Fundamental Class of Geodesics on any Closed Surface of Genus Greater than One," Trans. Am. Math. Soc. **26**, 25–60.
Moser, J., 1973, *Stable and Random Motions in Dynamical Systems* (Princeton University, Princeton, NJ).
Murray, N. W., 1991, "Critical Function for the Standard Map," Physica D **52**, 220–245.
Mynick, H. E., and J. A. Krommes, 1980, "Particle Stochasticity due to Magnetic Perturbations of Axisymmetric Geometries," Phys. Fluids **23**, 1229–1237.
Ottino, J. M., 1989, *The Kinematics of Mixing: Stretching, Chaos, and Transport* (Cambridge University, Cambridge, England).
Percival, I. C., 1979a, "A Variational Principle for Invariant Tori of Fixed Frequency," J. Phys. A **12**, L57–L60.
Percival, I. C., 1979b, "Variational Principles for Invariant Tori and Cantori," in *Nonlinear Dynamics and the Beam-Beam Interaction,* edited by M. Month and J. C. Herrera (AIP, New York), pp. 302–310.
Percival, I. C., 1982, "Chaotic Boundary of a Hamiltonian Map," Physica D **6**, 67–77.
Percival, I. C., and F. Vivaldi, 1987a, "Arithmetical Properties of Strongly Chaotic Motion," Physica D **25**, 105–130.
Percival, I. C., and F. Vivaldi, 1987b, "A Linear Code for the Sawtooth and Cat Maps," Physica D **27**, 373–386.
Pesin, Y. B., 1977, "Lyapunov Characteristic Exponents and the Smooth Ergodic Theory," Russ. Math. Surv. **32**, 55–114.
Petschel, G., and T. Geisel, 1991, "Unusual Manifold Structure and Anomalous Diffusion in a Hamiltonian Model for Chaotic Guiding Center Motion," Universität Frankfurt preprint.
Poincaré, H., 1885, "Mémoire sur les Courbes Définies par une Équation Différentielle, III," J. Math. Pures Appl. **1**, 167–244.
Poincaré, H., 1892, *Les Méthodes Nouvelles de la Mécanique*

Céleste (Gauthier-Villars, Paris).
Quispel, G. R. W., J. A. G. Roberts, and C. J. Thompson, 1989, "Integrable Mappings and Soliton Equations II," Physica D **34**, 183–192.
Rechester, A. B., and M. N. Rosenbluth, 1978, "Electron Heat Transport in a Tokamak with Destroyed Magnetic Surfaces," Phys. Rev. Lett. **40**, 38–41.
Rechester, A. B., M. N. Rosenbluth, and R. B. White, 1981, "Fourier-Space Paths Applied to the Calculation of Diffusion for the Chirikov-Taylor Model," Phys. Rev. A **23**, 2664–2672.
Rom-Kedar, V., 1990, "Transport Rates of a Class of Two-Dimensional Maps and Flows," Physica D **43**, 229–268.
Rom-Kedar, V., A. Leonard, and S. Wiggins, 1990, "An Analytical Study of Transport, Mixing, and Chaos in an Unsteady Vortical Flow," J. Fluid Mech. **214**, 347–394.
Rom-Kedar, V., and S. Wiggins, 1988, "Transport in Two-Dimensional Maps," Arch. Ration. Mech. Anal. **109**, 239–298.
Rosenbluth, M. N., R. Z. Sagdeev, J. B. Taylor, and G. M. Zaslavski, 1966, "Destruction of Magnetic Surfaces by Magnetic Field Irregularities," Nucl. Fusion **6**, 297–300.
Schmidt, G., and J. Bialek, 1982, "Fractal Diagrams for Hamiltonian Stochasticity," Physica D **5**, 397–404.
Sevryuk, M. B., 1986, *Reversible Systems,* Lecture Notes in Mathematics Vol. 1211 (Springer-Verlag, New York).
Skodje, R. T., and M. J. Davis, 1988, "A Phase Space Analysis of the Collinear $I+HI$ reaction," J. Chem. Phys. **88**, 2429–2456.
Stark, J., 1986, "On Invariant Circles for Area-Preserving Maps," Ph.D. thesis (University of Warwick).
Suris, Y. B., 1989, "Integrable Mappings of the Standard Type," Funct. Anal. Appl. **23**, 74–76.
Umberger, D. K., and J. D. Farmer, 1985, "Fat Fractals on the Energy Surface," Phys. Rev. Lett. **55**, 661–664.
Veerman, J. J. P., and F. M. Tangerman, 1991, "Intersection Properties of Invariant Manifolds in Certain Twist Maps," Commun. Math. Phys. **139**, 245–265.
Veselov, A. P., 1988, "Integrable Discrete-Time Systems and Difference Equations," Funct. Anal. Appl. **22**, 83–93.
Wigner, E., 1937, "Calculation of the Rate of Elementary Association Reactions," J. Chem. Phys. **5**, 720–725.
Wojtkowski, M., 1981, "A Model Problem with the Coexistence of Stochastic and Integrable Behaviour," Commun. Math. Phys. **80**, 453–464.
Zakharov, V. E., 1991, *What is Integrability?,* Nonlinear Dynamics (Springer-Verlag, Berlin).