LBL-27433

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

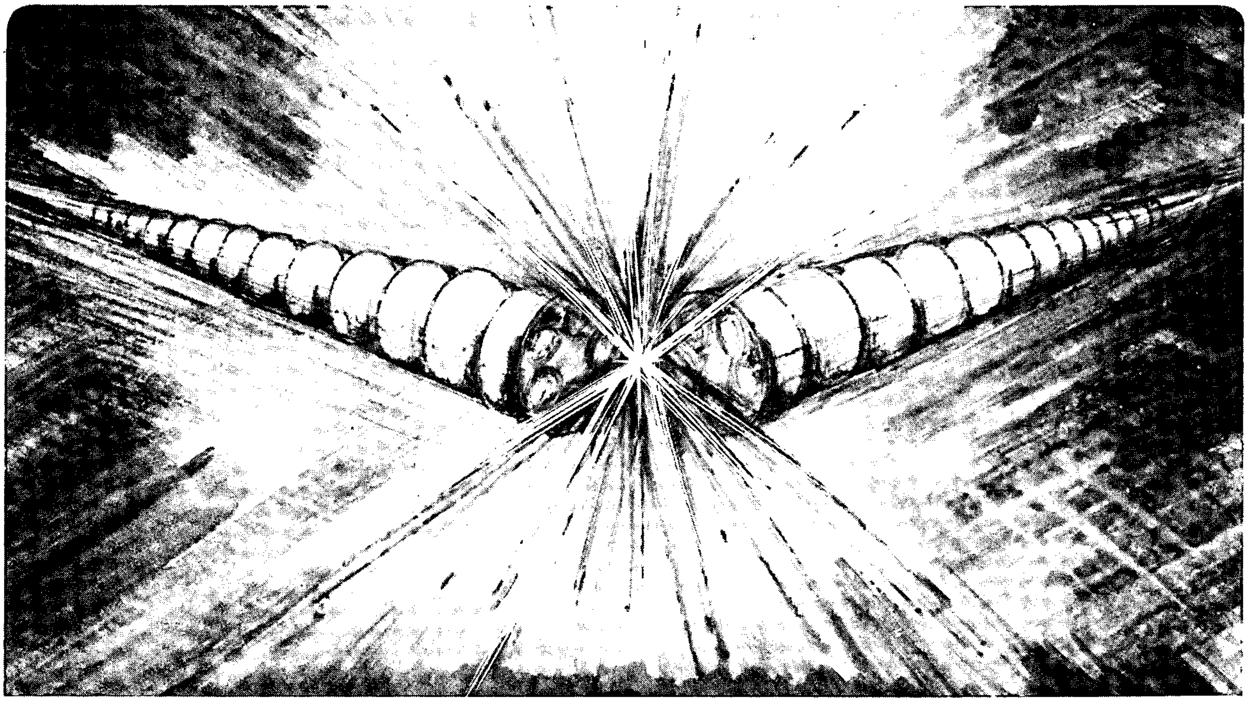
Accelerator & Fusion
Research Division

Lectures presented at the Dynamics Seminar, Berkeley, CA,
Spring Semester 1989, and to be published in *Review of Modern Physics*

Symplectic Maps, Variational Principles, and Transport

J. Meiss

April 1989



Prepared for the U.S. Department of Energy under Contract Number DE-AC03-76SF00098.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not nccessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

Symplectic Maps, Variational Principles, and Transport

J.D. Meiss
Accelerator and Fusion Research Division
Lawrence Berkeley Laboratory
1 Cyclotron Road
Berkeley, California 94720

April 1989

Contents

1 Symplectic Mappings................1
1 Return Mappings....................................1
A Example–Hénon-Heiles Hamiltonian...3
B Example–Volume Preserving Flow......4
2 Twist Mappings.......................................5
3 Examples...6
A The Cyclotron.....................................6
B Poincaré Section..................................7
C Incommensurate States........................8
D Billiards in a Convex Domain..............9
2 Orbits, Stability and Number Theory.................................11
1 Periodic Orbits......................................11
2 Stability..11
3 Stable Manifolds....................................12
4 Frequency..14
5 Number Theory......................................15
A Diophantine Numbers.........................15
B Continued Fractions..........................15
C Farey Tree..17
6 KAM Theory..19
3 Invariant Circles.......................21
1 Rotational Invariant Circles....................21
2 Net Flux...22
3 Birkhoff's Theorem.................................22
A Proof of Birkhoff's Theorem.............24
B Lipschitz Corollary............................26
C Confinement Corollary......................27
D Converse KAM theory......................28
4 Variational Principles................30
1 Generating Function...............................30
2 Net Flux...32
3 Examples...33
4 Action...34
5 Periodic Orbits.........................37
1 Minimizing orbits....................................37
2 Existence of (m,n) orbits.......................38
3 Aubry's Fundamental Lemma..................40
4 Minimizing (m,n) orbits..........................44
5 Minimax Principle..................................44
6 Quasiperiodic Orbits.................47
1 Invariant Circles are Minimizing.............47
2 Monotone Sets.......................................49
3 Existence of Quasiperiodic Orbits...........53
4 Cantori..54
5 Characterization of the set of Minimizing Orbits..56
6 Heteroclinic Orbits................................57
7 Mather's DW...57
7 Flux...58
1 Irregular Components............................58
2 Flux–Definition......................................59
3 The Important Question..........................60
4 Partial Barriers and Turnstiles.................61
5 Heteroclinic Orbits.................................63
6 Cantori..64
7 Resonances...64
8 Areas and Actions–the Fundamental Formula..65
9 Flux–Periodic Orbits..............................66
10 Stable and unstable segments..............66
11 Flux–Homoclinic Orbits........................69
12 Areas–Cantori and Separatrices...........69
13 Chen's Area Formulae..........................71
14 Area Devil's Staircase..........................72
Appendix A
Differential Forms.....................................73
Appendix B
Circle Maps...74
References................................77

Acknowledgements:

These notes are from a series of lectures I 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 Lawrence Berkeley Laboratory. 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 helped to improve the clarity of my presentation. My understanding of this subject has grown out of a six year, extremely fruitful collaboration with Ian Percival and Robert MacKay, and I used the lecture notes of MacKay and Stark (MacKay and Stark, 1985) heavily in preparing my lectures.

My intention here is first to make the important work of Aubry and Mather accessible to physicists interested in Hamiltonian systems, and second to review the application of this theory to transport. The reader will note several places where these notes are incomplete (e.g. §6.5), and also the absence of a promised chapter (§8) on transport models. My hope is to eventually remedy this, and to add discussions on the phenomenology of maps as revealed by computer experiment, as well as a discussion of numerical techniques for finding periodic orbits.

This work was supported by the US Department of Energy under contract DE-AC03-76SF00098.

1 Symplectic Mappings

Symplectic, or Hamiltonian mappings arise in many applications. Here we'll review a few examples, and then define a particular type of symplectic mapping, the twist mapping, which will be our major concern.†

1.1 Return Mappings

Symplectic mappings often arise from Hamiltonian systems as return mappings. Consider a Hamiltonian flow on a 2N+2 dimensional phase space, with coordinates $z_0 = (q_0, p_0), z_1, \dots z_N$. To fix ideas, we might have a system of interacting particles, defined by a Hamiltonian $H(z) = {}^1/_2 \, \Sigma \, p_i^2 + V(q)$. Configuration coordinates are denoted by q_i and their conjugate canonical momentum are denoted by p_i. The equations of motion are

$$\begin{aligned} \frac{dq_i}{dt} &= \frac{\partial H}{\partial p_i} \\ \frac{dp_i}{dt} &= -\frac{\partial H}{\partial q_i} \end{aligned} \quad \Rightarrow \quad \frac{dz}{dt} = J\,\nabla H$$

Here J, the Poisson matrix, is the (2N+2)×(2N+2) anti-symmetric matrix given by

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

in (q_i,p_i) coordinates.

Since the energy is conserved, the flow occurs on a 2N+1 dimensional energy surface $\mathcal{E}$ corresponding to a value E = H; assume this surface is bounded. Suppose there is another 2N+1 dimensional surface $\mathcal{Q}$ which is locally transverse to the flow. The Poincaré section, $\mathcal{S}$, is the 2N dimensional intersection of $\mathcal{E}$ with $\mathcal{Q}$. The return mapping is the function which takes an initial condition z on $\mathcal{S}$, and gives the point z′

† For an alternative viewpoint on some of the topics in this chapter consult [Lichtenberg and Lieberman, 1982) or for the mathematically inclined (Arnol'd, 1978).

on $\mathcal{S}$ to which the flow first returns. Since the energy surface is bounded, almost all trajectories which begin on $\mathcal{S}$ will return to $\mathcal{S}$.

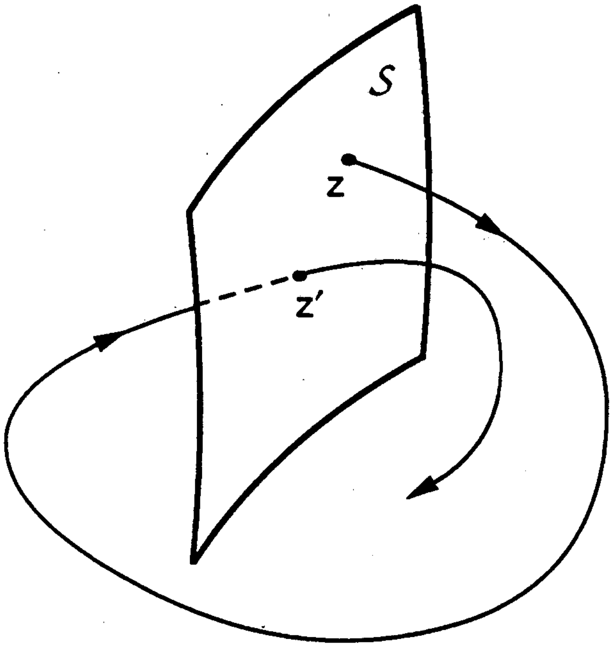


Figure 1.1.1 XBL 896-2158

For example, let $\mathcal{Q}$ be the surface q_0 = constant. It is transverse to the flow if $dq_0/dt = \partial H/\partial p_0 \neq 0$ on $\mathcal{Q}$. The Poincaré section $\mathcal{S}$ can be described by the coordinates $z_1,...z_N$, since with a choice of value for the energy, transversality implies that $H(q_0,p_0,z_1,...z_N) = E$ can be inverted to obtain $p_0 = p_0(z_1,...z_N; q_0, E)$ (often this function will have several branches: choose one). The return mapping, which we denote by T (it is parameterized by the choice of E and q_0) gives the next intersection point on $\mathcal{S}$ as a function of the initial coordinates $z_1, ...z_N$:

$$z' = T(z)$$

To show that the return mapping is symplectic, we obtain a reduced Hamiltonian description of the motion in terms of the function $p_0(z_1,...z_N; q_0, E)$. The equations of motion can be rewritten in the form:

$$\frac{dq_i}{dq_0} = \frac{\frac{dq_i}{dt}}{\frac{dq_0}{dt}} = \frac{\frac{\partial H}{\partial p_i}}{\frac{\partial H}{\partial p_0}} = \frac{\partial}{\partial p_i}[-p_0(z_1,\cdots,z_N;q_0,E)]$$

$$i = 1,\cdots N$$

$$\frac{dp_i}{dq_0} = \frac{\frac{dp_i}{dt}}{\frac{dq_0}{dt}} = \frac{-\frac{\partial H}{\partial q_i}}{\frac{\partial H}{\partial p_0}} = -\frac{\partial}{\partial q_i}[-p_0(z_1,\cdots,z_N;q_0,E)]$$

so that $-p_0$ acts as the Hamiltonian for the flow on the 2N dimensional phase space $\mathcal{S}$. The role of time is played by q_0; the Hamiltonian $-p_0$ is an explicit function of "time." In fact this reduction is precisely the inverse of the procedure for obtaining an "extended" phase space description of a time dependent Hamiltonian.

The mapping $z' = T(z)$ from $\mathcal{S}$ to $\mathcal{S}$ is obtained by following this Hamiltonian flow from a particular time q_0 to the next intersection with the surface $\mathcal{S}$, when q_0 returns to its initial value. Such a Hamiltonian flow preserves the Poincaré integral invariant, $\oint p\cdot dq$, where the loop integral is taken over any closed loop. A local statement of this is that the differential form (see appendix A) $\omega = \Sigma dp_i \wedge dq_i$ (the symplectic form) is preserved by the flow; therefore the mapping must preserve this form as well. This simply means that the Jacobian matrix corresponding to the derivative of the mapping, $M = \partial T(z)/\partial z$, is a symplectic matrix; that is, it preserves the canonical matrix J under congruency:

$$\tilde{M}JM = J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}$$

where J is now the 2N×2N antisymmetric matrix shown and ~ indicates transpose.

In the particular case of a two degree of freedom Hamiltonian, N=1, the mapping T acts on the two dimensional phase space (q_1,p_1). The symplectic condition implies that det(M) = 1 (note that it cannot be -1), or equivalently that the mapping preserves the oriented area element $dp_1 \wedge dq_1$.

A Example–Hénon-Heiles Hamiltonian

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

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

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 $E \leq 1/6$. 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 new "Hamiltonian" $-p_x(y,p_y;E,x)$. Since $p_x^2 \geq 0$, the domain of the mapping is restricted to the region $p_y^2 + y^2 - {}^2/_3 y^3 \leq E$, which looks like an oval for E small, and has a corner when E = 1/6.

B Example–Volume Preserving Flow

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. For example consider an incompressible fluid with velocity field v(x), or a magnetic field B(x). The equations for the Lagrangian particle trajectories, or field lines, are

$$\frac{dx}{dt} = B(x)$$

where t parameterizes the field lines. This system can be thought of as Hamiltonian in the following way (Cary and Littlejohn, 1983). Suppose the field we are describing has a non-vanishing component corresponding to some angle; call this direction ζ, the toroidal direction. By a choice of gauge the vector potential can be written $A = \varphi\nabla\zeta - \psi\nabla\theta$, where θ is another angle, the poloidal angle. The field is then $B = \nabla\varphi\times\nabla\zeta - \nabla\psi\times\nabla\theta$. We have assumed that the contravariant component of 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 implies that ψ is a good "radial" coordinate. Thus we use (ψ,θ,ζ) as coordinates and think of φ as a function, $\varphi(\psi,\theta,\zeta)$, of the other three coordinates. The field line equations become

$$\frac{d\theta}{d\zeta} = \frac{B^\theta}{B^\zeta} = \frac{\partial\varphi}{\partial\psi}, \qquad \frac{d\psi}{d\zeta} = \frac{B^\psi}{B^\zeta} = -\frac{\partial\varphi}{\partial\theta}$$

which are Hamilton's equations with the Hamiltonian φ, canonical variables (ψ,θ) and time ζ. Periodicity in ζ implies that we can use the Poincaré section technique to construct an area preserving mapping T: $(\psi,\theta) \to (\psi',\theta')$

1.2 Twist Mappings

Consider a phase space (x,y) which is a cylinder, $\mathcal{T}^1 \times \mathcal{R}^1$†, with x being the angle coordinate. Let T: $(x,y) \to (x',y')$ be an area preserving mapping from the cylinder to itself, and suppose T is differentiable ($\mathcal{C}^1$). Then T is a twist mapping (with twist to the right) if

$$\left.\frac{dx'}{dy}\right|_x \geq K > 0$$

Geometrically this means that the first iterate of a vertical line (x=constant) tilts to the right (is a graph over x).

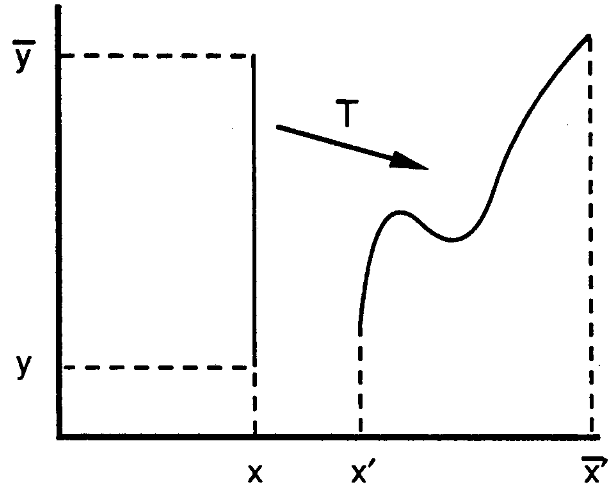


Figure 1.2.1 XBL 896-2119

This relation does not imply that y′(x,y) is a function of y.

Since the mapping is differentiable, we can consider its action on a tangent vector (δx,δy):

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

We will usually denote the linearized mapping by M. A symplectic mapping preserves the differential form $dy \wedge dx$, and so in two dimensions it preserves area and orientation. Thus the matrix M has unit determinant: det M = 1. The inverse of the linear mapping is represented by the derivative of T^{-1} as well as the inverse of M, thus

†We denote the 1-dimensional torus by $\mathcal{T}^1$ and the real line by $\mathcal{R}^1$.

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

Therefore the twist condition implies that

$$\left.\frac{\partial x}{\partial y'}\right|_{x'} = -\left.\frac{\partial x'}{\partial y}\right|_{x} \leq -K$$

so if T is a twist mapping, then T^{-1} is also a twist mapping, but twists to the left. Note that T^2 is not necessarily a twist mapping, 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 mappings, called "tilt" mappings, to which we will refer in §3.3).

This paper will almost entirely concentrate on the study of area-preserving twist mappings, firstly because the theory is well developed, and the twist condition permits the proof of several important theorems, secondly because twist mappings occur commonly in applications.

1.3 Examples

A The Cyclotron

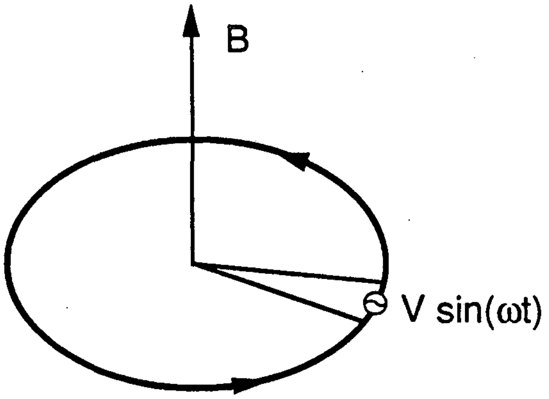


Figure 1.3.1 XBL 896-2118

Suppose there is a time dependent voltage drop Vsinωt across a narrow azimuthal gap in a magnetic field $B = B_0\, e_z$. The time for an electron to go around one circuit of the cyclotron is

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

where E is the particle energy $m\gamma c^2$. 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

$$E' = E - eV\sin\omega t \qquad t' = t + (2\pi/ceB)E'$$

(providing the kick is too small to reverse the velocity). By convention we define normalized variables $y = \omega/\Omega_c$, $x = \omega t/2\pi$, and $k = 2\pi\omega V/cB$. Then we get exactly the "standard map"

$$T: \begin{array}{l} y' = y - \frac{k}{2\pi}\sin(2\pi x) \\ x' = x + y' \end{array}$$

It depends on a single parameter, k representing the strength of the nonlinear kick. In the second equation, y′ must be taken to be the given function of (x,y) in order to obtain area preservation. This mapping has twist, in fact $\partial x'/\partial y = 1$.

B Poincaré Section near an Elliptic periodic orbit

Consider a two degree of freedom system. Take the Poincaré section to be transverse to some elliptic periodic orbit.

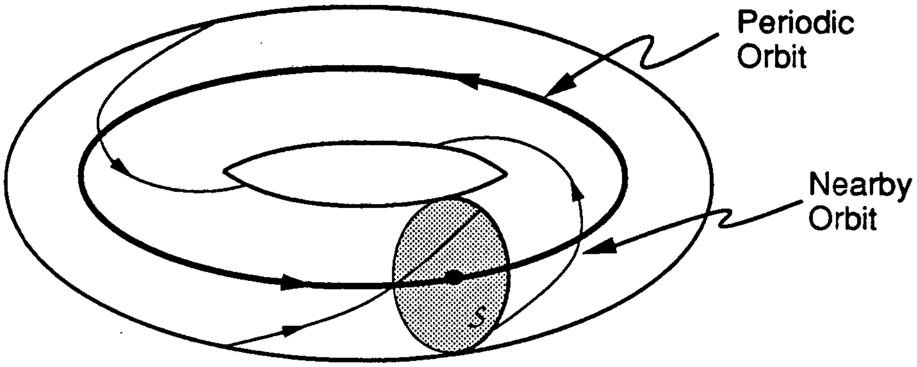


Figure 1.3.2 XBL 896-2120

By definition, an orbit is elliptic if its Floquet multipliers are on the unit circle; in other words, the return mapping T, has a linearization M with eigenvalues $e^{\pm 2\pi i\omega}$. Suppose we express T in generalized polar coordinates (r,θ) about the fixed point. Birkhoff shows that when ω is irrational a formal perturbation expansion for the mapping near the fixed point can be obtained in the form (see appendix 7 of (Arnol'd, 1978))

$$T: \begin{array}{l} r' = r + h(r,\theta) \\ \theta' = \theta + 2\pi\omega + \rho_2 r^2 + \cdots + \rho_{2k} r^{2k} + g(r,\theta) \end{array}$$

where h and g are $\mathcal{O}(r^{2k})$, and k can be made is as large as one likes. The mapping preserves the area rdrdθ. If any of the ρ_{2n} are not zero, then the mapping has twist providing r small enough. Thus in the neighborhood of an elliptic orbit, the flow tends to resemble a family of nested tori. In fact this formal power series does not converge, and the nested tori do not exist in general (see the discussion of the KAM theorem in §2.6).

C Incommensurate States

Consider a one dimensional chain of particles connected by harmonic springs. For simplicity, take the spring constants to be one.

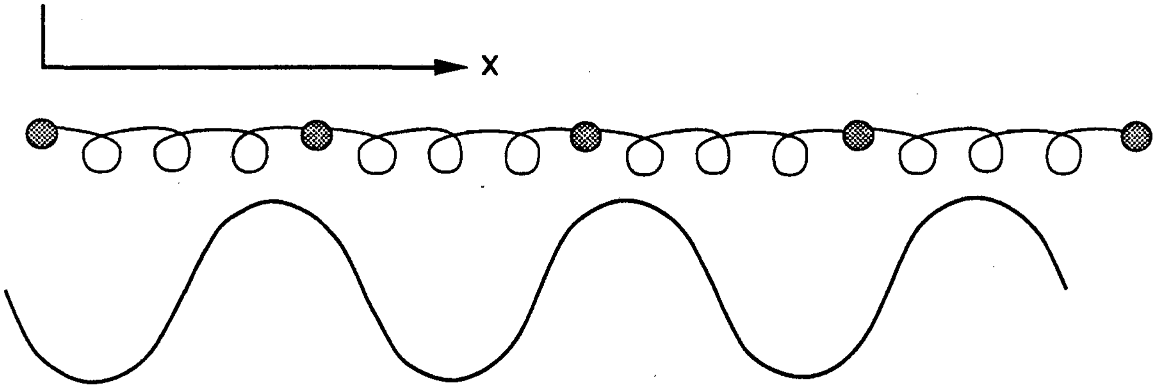


Figure 1.3.3 XBL 896-2121

Put these particles in a periodic potential V(x) = k/4π^2 cos(2πx). Equilibrium states are governed by force balance:

$$\left(x_{j+1} - x_j\right) - \left(x_j - x_{j-1}\right) + \frac{k}{2\pi}\sin(2\pi x_j) = 0$$

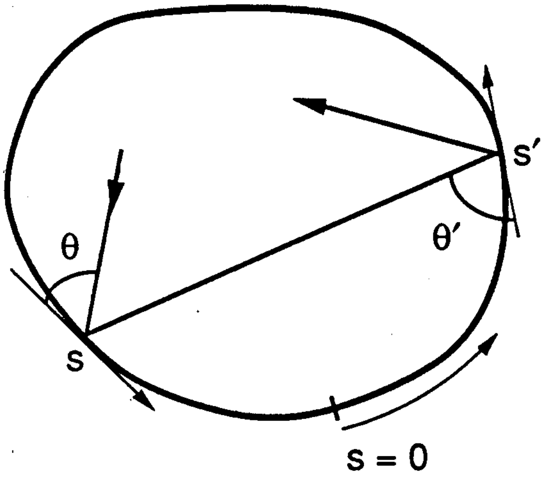
If we define $y_j = x_j - x_{j-1}$, then this becomes the standard map, where "time" is reinterpreted as the particle index j. This model is known as the Frenkel-Kontorova model (Aubry, 1983). Aubry studied the nature of the ground states of this system. These can be loosely defined as states of minimum energy, where the energy of a configuration is

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

We will learn much about this function and its extrema later.

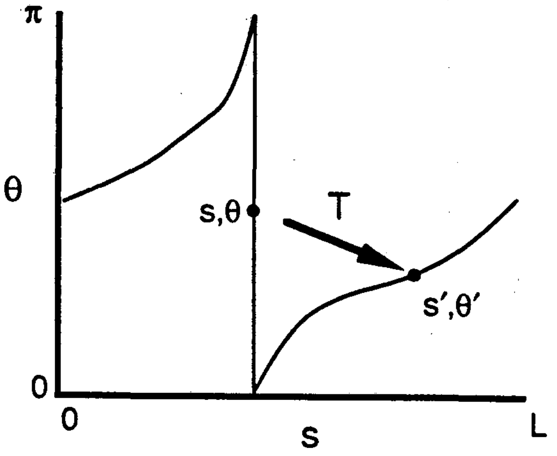
D Billiards in a Convex Domain (Berry, 1981)

Consider an particle bouncing elastically in a two dimensional, convex domain. A twist mapping for this system is determined by the sequence of boundary points at which the bounces occur. Following Birkhoff , we use the arc length along the boundary from a given point, s∈[0,L], and the angle between a tangent to the boundary and the trajectory, θ∈[0,π], for coordinates. It is easy to see that s′(s,θ) is a monotone increasing function of θ because of the convexity of the boundary.



XBL 896-2151

Figure 1.3.4



XBL 896-2152

Figure 1.3.5

In fact s is an angle-like coordinate since the mapping is periodic with period L in s. As we will see in §4.3 this mapping preserves the measure sinθ dsdθ, so a good set of coordinates to use is (x,y) = (s, cosθ). The boundaries y = ± 1 are fixed points (unfortunately the twist dx′/dy vanishes at these points).

2 Orbits, Stability and Number Theory

2.1 Periodic Orbits

Consider an area preserving mapping T on the cylinder. An orbit is a sequence of phase space points (x_i, y_i) for all i. To define periodic orbits it is convenient to "lift" the angle coordinates to the real line. For example, for the standard mapping this corresponds to computing $x' = x + y'$ without taking the fractional part. Such a lift is unique up to an integer shift in x. When lifted the mapping takes the plane $\mathcal{R}^2$ to itself.

An orbit is periodic with period n if n is the smallest integer such that

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

for some integer m. We will denote such an orbit by (m,n). We will see that there are at least two such orbits for a twist mapping: roughly speaking one forms the center of an island chain, and the other is the hyperbolic orbit which gives gives rise to the separatrix.

2.2 Stability

Stability of an orbit is measured by the properties of the tangent mapping:

$$\delta z_n = \left(\frac{d}{dz_0} T(T(\cdots T(z_0))) \right) \delta z_0 = M \, \delta z_0$$

Here M denotes the 2×2 Jacobian matrix obtained from the derivative of T^n. There are two invariant properties associated with M, its determinant and its trace; since T is area preserving det(M) = 1. Thus M has two eigenvalues λ and $1/\lambda$ which are the solutions of the characteristic polynomial

$$\lambda^2 - \mathrm{Tr}(M)\,\lambda + 1 = 0 \rightarrow \lambda = \frac{1}{2}\left(\mathrm{Tr}(M) \pm \sqrt{[\mathrm{Tr}(M)]^2 - 4}\right)$$

A stability classification is most conveniently given in term of the residue (Greene, 1979):

$$R = {}^1/_4[2 - \mathrm{Tr}(M)]$$

The possible stability properties are

a) Hyperbolic: both eigenvalues are real and positive.
b) Elliptic: there is a pair of complex conjugate eigenvalues with unit modulus.
c) Reflection Hyperbolic: both eigenvalues are real and negative.
d) Parabolic: The eigenvalues are both 1 or both -1.

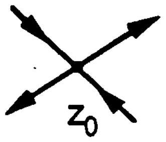
Stability	λ	R	Tr(M)
hyperbolic	> 0	< 0	> 2
elliptic	$e^{2\pi i\omega}$	(0,1)	(-2,2)
reflection hyperbolic	< 0	> 1	< -2

The elliptic case is the only one which could possibly be called stable, although the stability is a neutral one. Linear (or spectral) stability does not guarantee that the orbit is actually stable (that is points initially close stay nearby); much more analysis is required, and the full apparatus of the KAM theorem must be used (Arnol'd, 1978). Positive residue corresponds either to an elliptic or reflection hyperbolic orbit. We will see later that these two cases are properly though 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 loose stability.

2.3 Stable Manifolds

For a hyperbolic period n orbit, the linear mapping has two eigenvectors corresponding to the unstable and stable directions ($\lambda>1$ and $1/\lambda < 1$, respectively). The stable manifold theorem (Lanford, 1973) implies that the eigenvectors can be extended to invariant manifolds W^u and W^s. Each point on these accumulates on the hyperbolic orbit in at least one direction of time:

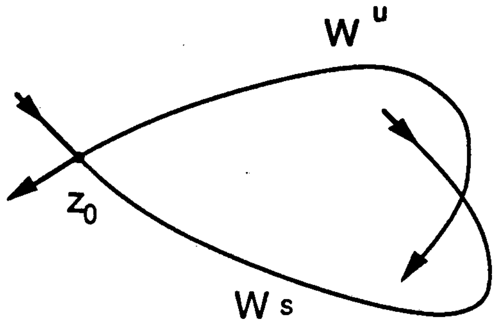
$$z \in W^s \Rightarrow T^{jn}z \rightarrow z_0 \text{ as } j \rightarrow \infty$$
$$z \in W^u \Rightarrow T^{jn}z \rightarrow z_0 \text{ as } j \rightarrow -\infty$$



XBL 896-2156

Figure 2.3.1

where z_0 is some point on the orbit. W^s (W^u) can not intersect itself or the stable (unstable) manifold of any other periodic orbit, since this would violate uniqueness. Generically W^u and W^s are different manifolds; one exception to this is an the integrable system for which W^u and W^s join smoothly to form a separatrix. Typically W^u and W^s intersect transversely (if at all); the intersections are called homoclinic points.

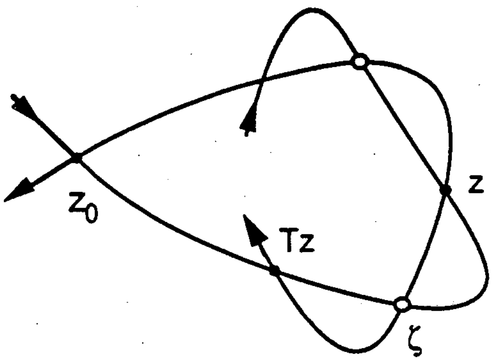


XBL 896-2154

Figure 2.3.2

Heteroclinic points are the intersection points of the stable and unstable manifolds of different periodic orbits.

The existence of a single homoclinic point implies that of an infinite number, because the crossing point lies both on W^u and W^s.



XBL 896-2153

Figure 2.3.3

Thus each iterate of a homoclinic point is also homoclinic. There also must be a second homoclinic point ζ between a homoclinic point z and T(z), since the mapping preserves orientation. Furthermore, Poincaré realized that there are an infinity of other homoclinic points which are not related by iteration of the mapping. This is because the single crossing implies that the homoclinic loops of W^u and W^s must curl around in an intricate way in order to attempt to remain in the region bounded by the curves W^u from $z_0 \rightarrow z$ and W^s from $z \rightarrow z_0$. Because of area preservation, this attempt ultimately fails, generating new intersections.

We will see in §6.6 that for a twist mapping, a theorem of Aubry and Le Daeron (Aubry and Le Daeron, 1983) implies that neighboring points on a hyperbolic (m,n) orbit necessarily have heteroclinic connections.

2.4 Frequency

To define the frequency of an orbit, we use the lift of the mapping to $\mathcal{R}^2$, and compute the limit

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

For a periodic (m,n) orbit, the frequency always exists, and is given by m/n. An orbit is called quasiperiodic if the frequency exists and is irrational, and the orbit is recurrent (it returns arbitrarily closely to each point on the orbit).

2.5 Number Theory

For small perturbations from an integrable twist mapping (e.g. $k = 0$ for the standard map), there appear to be many quasiperiodic orbits. Their persistence depends on the fact that some irrational numbers are "far" from rationals.

A 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. In particular, an irrational is particularly hard to approximate if it satisfies a Diophantine condition: there exists a $C>0$ such that for all integers (m,n), there is a $\tau \geq 1$ such that

$$|n\omega - m| > C/n^{\tau}$$

For C small enough the set of such numbers is not empty; in fact for any $\tau > 1$ the measure of ω satisfying a Diophantine condition approaches one as C approaches zero.

B Continued Fractions

Another method for classification of the properties of real numbers is the continued fraction (Khinchin, 1964). The continued fraction of ω is the sequence $[a_0, a_1, ...]$ of integers generated by the mapping

$$\begin{aligned} a &= [\omega] \\ \omega' &= 1/(\omega - a) \end{aligned}$$

where the square brackets indicate the nearest integer less than ω (If ω is negative, a_0 negative; the remaining a_i are positive). An alternative representation for the continued fraction is

$$\omega = a_0 + {}^{1}\!/\!(a_1 + {}^{1}\!/\!(a_2 + \cdot\cdot 1/(a_n + \cdots$$

The continued fraction expansion of an irrational is infinite, while that for rationals always ends (one eventually finds that ω' is an integer). Every rational has two equivalent continued fraction representations:

$$[a_0, a_1, \ldots, a_i] = [a_0, a_1, \ldots, a_i - 1, 1],$$

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

$$m_i/n_i = [a_0, a_1, \ldots, a_i]$$

The continued fraction expansion is a strongly convergent expansion: for any ε there is a j such that the norm

$$|m_i\omega - n_i| < \varepsilon \quad \forall \; i \geq j.$$

Furthermore the convergents are best approximants: for a convergent m/n to ω there are no m'/n' with $n' \leq n$ satisfying $|m'\omega - n'| < |m\omega - n|$.

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

$$|m\omega - n| < C/n \qquad (*)$$

for $C=1$; furthermore, every rational which satisfies this inequality when $C = 1/2$ is a convergent. However, when C is too small ($C < 1/\sqrt{5}$), there exist ω such that only finitely many convergents satisfy the inequality.

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

$$\pi = [3,7,15,1,292,1,1,1,2,1\ldots]$$

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 $a_i < \alpha$, $\forall i$. For such ω, and for sufficiently small C, there are no (m,n) satisfying the inequality (*). In fact the numbers of constant type are precisely those which satisfy a Diophantine condition 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. Liouville showed that every quadratic irrational has an eventually periodic continued fraction, and conversely every 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. Interestingly, not much is known about these when the degree is larger than two, except that any algebraic number satisfies a Diophantine condition with any $\tau>1$.

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:

$$\gamma = \frac{1+\sqrt{5}}{2} = [1,1,1,....]$$

which is naturally also a quadratic irrational.

C Farey Tree

The Farey tree is a technique for organizing the rational numbers according to the length of their continued fraction expansions. The tree is constructed by beginning with a pair of rationals, m/n and m′/n′, which are neighboring: mn′ - nm′ = 1 and which are written in lowest common terms. Level one of the tree is generated from these two by by adding their numerators and denominators

$$\frac{m''}{n''}=\frac{m+m'}{n+n'}$$

This rational is the mediant of m/n and m′/n′. It is not difficult to see that m″ and n″ have no common factors, 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 which gives every rational number in the interval [m′/n′, m/n]. The master tree, which gives all the positive numbers is generated by [0/1,1/0]:

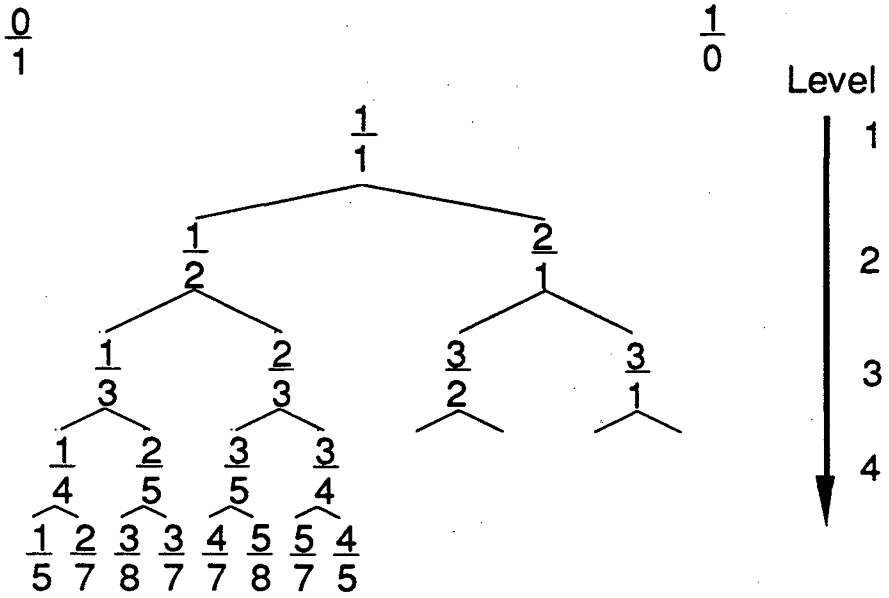


Figure 2.5.1 XBL 896-2139

The Farey path for a number is the sequence of steps leading to it from 1/1. Irrationals are represented by infinitely long Farey paths. In fact the Farey path provides a binary code for the reals: assign a symbol (L or R) to each step according to whether the step is to the left or right.

The continued fraction expansion is closely related to the Farey tree construction. The sum of the continued fraction elements of m/n gives the level which it occurs in the tree:

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

To obtain the continued fraction for a rational on the tree follow the Farey path leading to it from 1/1 = [1] = [0,1]. The rule to go from a rational m/n to one on the next level is to increment the last continued fraction element of m/n by one; use the representation of m/n with $a_i \neq 1$ if the current step in the Farey path is in the same direction as the preceding step, otherwise use the representation with last element equal to one:

$$\frac{m}{n} = \begin{Bmatrix} [a_0, a_1, \cdots a_i] \\ [a_0, a_1, \cdots a_i - 1, 1] \end{Bmatrix} \Rightarrow \begin{Bmatrix} [a_0, a_1, \cdots a_i + 1] & \leftarrow \text{no direction change} \\ [a_0, a_1, \cdots a_i - 1, 2] & \leftarrow \text{direction change} \end{Bmatrix}$$

For example the golden mean corresponds to the path RLRLR.... = [1,1,1,1,1,1...]. In general, noble numbers have a Farey path which eventually becomes an alternating sequence, ...LRLR... .

There are two important types of infinite Farey paths: those which eventually consist of all L's or all R's, and those which continue to alternate. The former paths converge to rational numbers either from above or below. For example the sequence RLLLLLLL... → 1/1 from above and LRRRRRRR...→ 1/1 from below. We will call these numbers $1/1|_{+}$ and $1/1|_{-}$. They are distinct from 1/1 and have a nice interpretation in terms of the orbits of a twist mapping. Farey paths which never settle down to one direction or the other approach irrational numbers.

2.6 KAM Theory

Consider an integrable area preserving twist mapping, e.g.

$$T_0 : \begin{cases} y' = y \\ x' = x + \Omega(y') \end{cases} \qquad d\Omega/dy > 1$$

The twist condition implies that there are quasiperiodic orbits for all irrational ω, in fact, since y is a constant of motion, the frequency is just $\omega = \Omega(y)$. Each quasiperiodic orbit densely covers a circle y = constant. In general, an invariant curve which is topologically equivalent to y=constant is called a rotational invariant circle (RIC).

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, we define the j-norm of a function f as

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

One version of the KAM theorem is

<u>Theorem</u> (Moser, 1973): For $\tau \geq 1$, $j > 2\tau+1$, and $\Omega(y) \in \mathcal{C}^j(\mathcal{R})$, there is an $\varepsilon>0$ such that for any $C>0$ all area preserving maps T such that $|T-T_0|_j < \varepsilon C^2$ on some annulus $y_0 < y < y_1$ have rotational invariant circles for all ω in an interval slightly smaller than $[\Omega(y_0), \Omega(y_1)]$ which satisfy a Diophantine condition for C and τ .

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

Most invariant circles persist for sufficiently small perturbations, however in the proof of the theorem "small" is indeed very small. The most highly optimized version of the KAM theorem is that of Herman (Herman, 1985), who showed that there is at least one invariant circle (with $\omega = \gamma$) of the standard map when $k \le 1/34$, $\tau = 1$, and $j = 4$. As we will see in the next chapter, it is often more efficient to ask the converse question: when do rotational invariant circles do not exist?

3 Invariant Circles

In this chapter we discuss the theorem of Birkhoff which implies that rotational invariant circles of a twist mapping must be Lipschitz graphs. This result has important applications to the theory of existence of invariant circles, and to transport.

3.1 Rotational Invariant Circles

Let T be an area preserving mapping on the cylinder $\mathcal{T}^1\times\mathcal{R}$ 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 mapping arises from a Poincaré section of a flow, since the flow provides a smooth connection of the mapping to the identity mapping.

An invariant circle is a curve $\mathcal{C}$ such that $T\mathcal{C} = \mathcal{C}$. A rotational invariant circle (RIC) is a homotopically non-trivial circle (it cannot be continuously deformed into a point). An invariant circle divides the cylinder into two invariant regions: to see this, consider the iterate of the region below an RIC; because the mapping is continuous, this iterate is a connected region, and because points far below remain far below, the mapping is one to one, and the circle is invariant, the iterate cannot cross the circle. Thus an invariant circle provides an absolute barrier to motion.

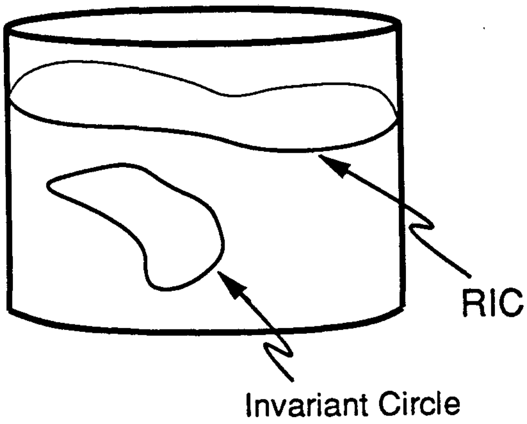


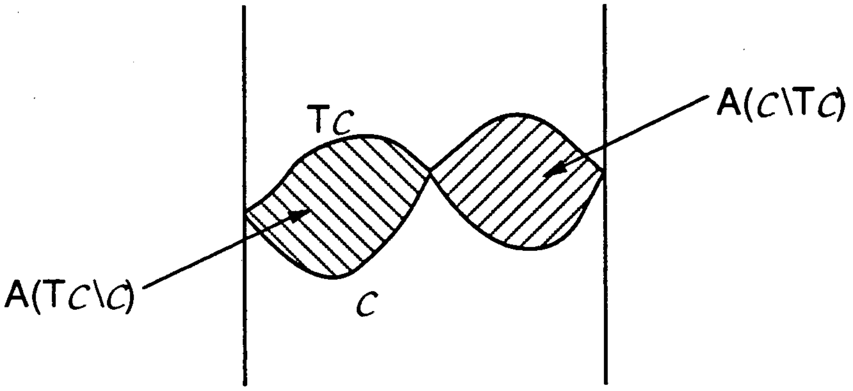
Figure 3.1.1 XBL 896-2138

3.2 Net Flux

Let T be an area preserving, end preserving mapping on the cylinder. Consider a loop $\mathcal{C}$ which encircles the cylinder once. The net flux is defined to be the area contained between $\mathcal{C}$ and T$\mathcal{C}$:

$$\mathcal{F} = A(T\mathcal{C}\backslash\mathcal{C}) - A(\mathcal{C}\backslash T\mathcal{C})$$

where $A(\mathcal{C}\backslash\, \mathcal{D})$ means the area of the region below $\mathcal{C}$ which is also above $\mathcal{D}$.



T$\mathcal{C}$
A($\mathcal{C}$\T$\mathcal{C}$)
A(T$\mathcal{C}$\$\mathcal{C}$)
$\mathcal{C}$

Figure 3.2.1 XBL 896-2137

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 them is invariant $A(\mathcal{C}\backslash\mathcal{D}) = A(T\mathcal{C}\backslash T\mathcal{D})$. Furthermore, the difference between these two areas is just the difference between the net flux through $\mathcal{C}$ and that through $\mathcal{C}'$.

A mapping with zero net flux is also called exactly symplectic (see also §4.2).

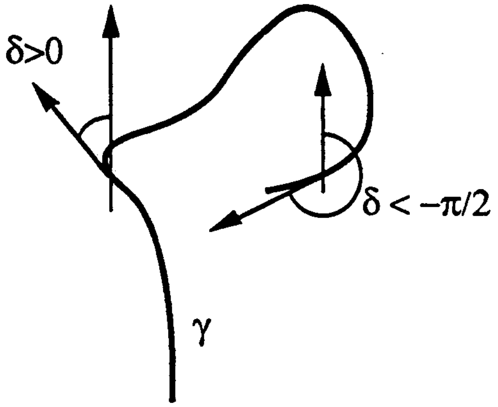
A mapping which has an RIC must have zero net flux, since the net flux through the RIC is zero.

3.3 Birkhoff's Theorem (Birkhoff, 1920; Mather, 1984)

Theorem: Suppose T is a C^1 area preserving, end preserving twist mapping on the cylinder. Let U be an open invariant set homeomorphic to the cylinder s.t. $\exists$ 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, by assumption can have no holes and must look like "half a cylinder " since it has an upper boundary contained between y=a and y=b. The important point of the theorem is that ∂U cannot have any "whorls", for example like those of a breaking wave. In particular, any rotational invariant circle can be used to form ∂U, so the theorem implies that all RIC's are graphs.

To prove this we introduce the notion of accessible points. Let $\gamma(t) = (x(t),y(t))$ be a curve embedded in U (γ cannot cross itself) parameterized by t, such that $y(t) \to -\infty$ as $t \to -\infty$. The deviation of γ from the vertical is defined to be the angle δ between a tangent to γ and the vertical.



XBL 896-2144

Figure 3.3.1

For for those points $\gamma(t)$ such that $y(t) > y(t')$ for all $t' < t$, choose δ in $[-\pi/2,\pi/2]$; otherwise the branch of δ is chosen to make the deviation a continuous function.

A curve γ is tilted to the right (left) if its deviation from the vertical is everywhere to the right (left), i.e. $\delta<0$ ($\delta>0$); such curves are denoted γ^R (γ^L).

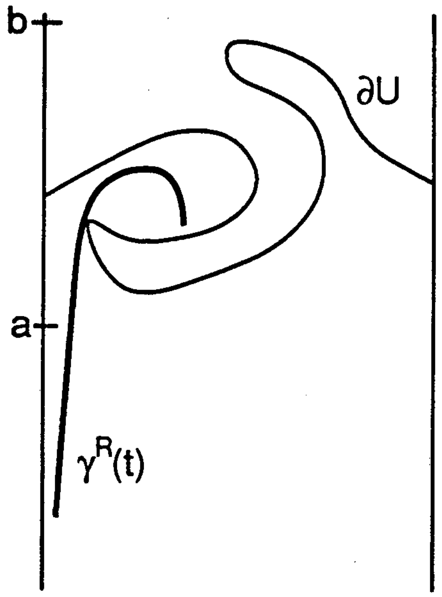


Figure 3.3.2 XBL 896-2145

A point $z_0 \in U$ is <u>right accessible</u> if there exists a $\gamma^R \in U$ such that $\gamma^R(t_0) = z_0$.

A Proof of Birkhoff's Theorem

A curve γ^R which tilts to the right is mapped onto another such curve by T. To see this let the angle between the vertical and a tangent to γ^R at z be $-\pi<\delta<0$. By the twist condition, the vertical vector v at z is mapped into a right tilting vector DT(v) with tilt $-\pi<\theta<0$. Since T preserves orientation, the angle δ' between DT(v) and the tangent to $T(\gamma^R)$ at T(z) must be in the range $[-\pi,0]$. The deviation of $T(\gamma^R)$ from the vertical is the sum of these two angles, and therefore must be to the right.

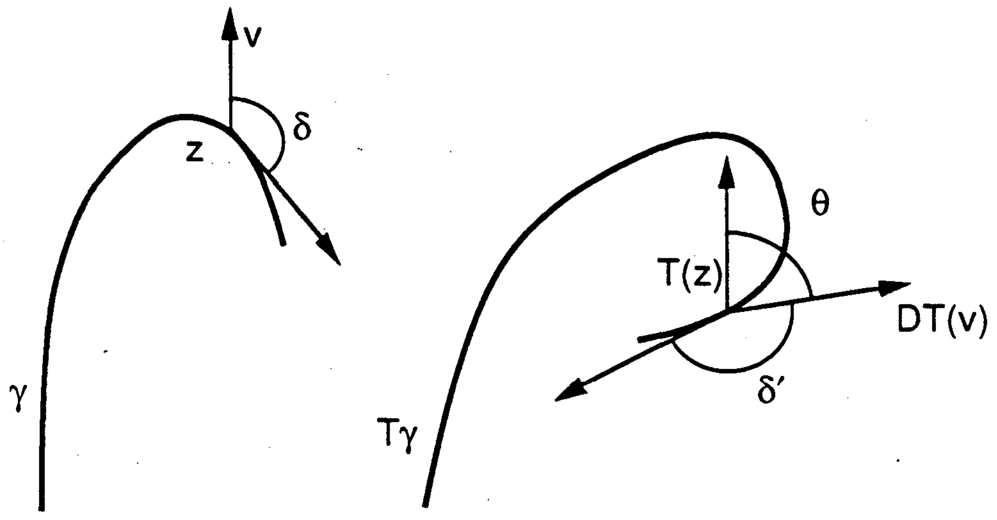


Figure 3.3.3 XBL 896-2146

Let W^R and W^L be the subsets of U which are right and left accessible, respectively. Note that ∂W^R consists of portions of ∂U together with vertical segments bounding those parts of U not right accessible.

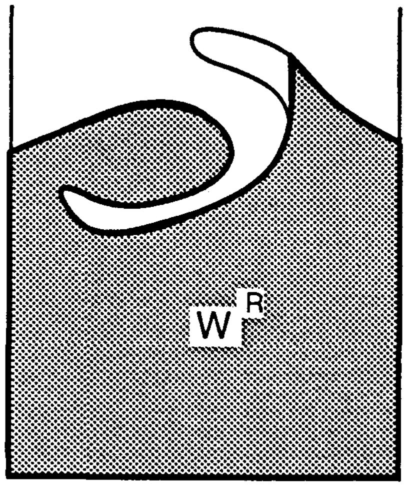


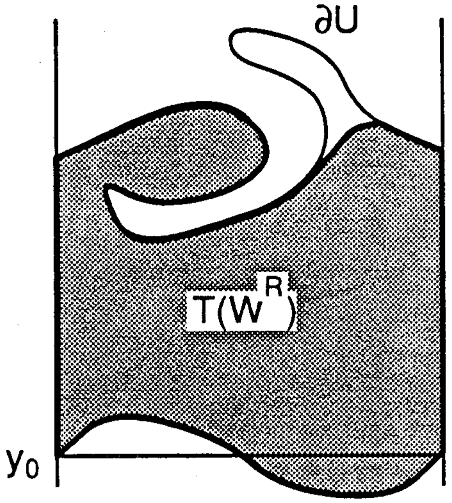
Figure 3.3.4 XBL 896-2149

Since every point in W^R is on a curve which 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 $W^R = U$. For suppose this is not true, then there is some portion of U which 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. However, this would violate area preservation because, consider a circle far below ∂U, for example $y = y_0$ where $y_0 << a$. Since ∂U is contained between y=a and y=b, the area of U above y_0 is finite. Furthermore, the area of W^R above $y = y_0$ is is mapped into a region with the same area. However, since the net flux through $y=y_0$ is zero, this gives a contradiction.



XBL 896-2150

Figure 3.3.5

Consideration of T^{-1} implies $W^L = U$.

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 ❑

Birkhoff's theorem has several important corollaries:

B Lipschitz Corollary

Any rotational invariant circle is a Lipschitz graph.

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

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

for all x_1 and x_0. These constants give a Lipschitz cone which contains the graph of the function. A Lipschitz function is continuous, and differentiable almost everywhere.

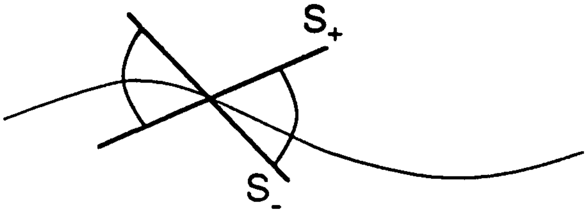


Figure 3.3.6 XBL 896-2148

Proof: For a twist mapping we can obtain explicit bounds on the slopes of an RIC. Upon iteration a vertical vector acquires a slope

$$S = \left.\frac{\delta y'}{\delta x'}\right|_x = \frac{\partial y'}{\partial y} \Big/ \frac{\partial x'}{\partial y}$$

Now the denominator is bounded below by the twist constant K, let S_+ be the maximum of this expression. Similarly by inverse iteration of a vertical let S_- be

$$S_- = \min\left[\frac{\partial y}{\partial y'} \Big/ \frac{\partial x}{\partial y'}\right] = \min\left[-\frac{\partial x'}{\partial x} \Big/ \frac{\partial x'}{\partial y}\right]$$

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 ❑

C 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.

Proof: Construct the region U for application of Birkhoff's theorem as follows: Let V be the invariant set formed from the iterates of all points y<a. This is not necessarily homotopic to the cylinder (there will typically be lots of holes in the annulus a<y<b corresponding to elliptic island chains). Let W be the complement of V. There is a

connected component of W which contains all points y>b. Let U be the complement of this component..U satisfies the hypotheses of Birkhoff's theorem ❑

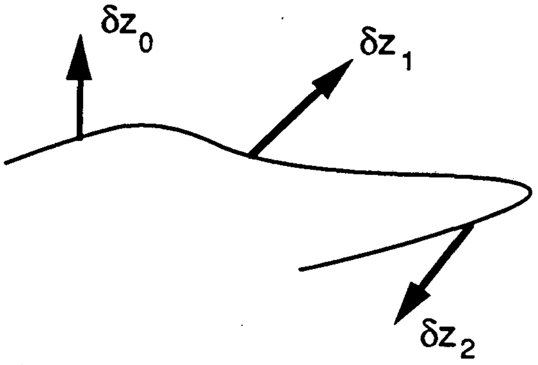
D Converse KAM theory: Non-existence of Rotational Invariant Circles

Birkhoff's theorem leads to several criteria for the non-existence of invariant circles, which are more or less effective techniques in practice.

1) Climbing Orbits: If there is an orbit which climbs arbitrarily far up the cylinder, then there are no invariant circles. More precisely consider an annulus $\mathcal{T}^1\times [a,b]$. 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 $2S_+|S_-|/(S_++|S_-|)$, 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 can take an extremely long time for orbits to climb even a small distance.

2) 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 is really what underlies the resonance overlap criterion of Chirikov; see e.g. (Lichtenberg and Lieberman, 1982); which is a perturbative attempt to find a heteroclinic intersection. One could follow these manifolds numerically.

3) Lipschitz Criteria: Using the Lipschitz bounds on slopes one can construct criteria for the non-existence of RIC's. Consider the iteration of a vertical unit vector $\delta z_0 = (0,1)$.



XBL 896-2147

Figure 3.3.7

Upon one iteration δz_0 becomes $\delta z_1 = (\partial x'/\partial y, \partial y'/\partial y)$, which has positive δx_1 by the twist condition. However, one more iteration can give a vector with $\delta x_2 < 0$, or

$$0 > \frac{\partial x''}{\partial x'}\frac{\partial x'}{\partial y} + \frac{\partial x''}{\partial y'}\frac{\partial y'}{\partial x} = 2 - k\cos(2\pi x')$$

where the last expression is that for the standard map (see §1.3.1). Since an RIC must intersect every vertical, if the inequality is satisfied for any x' there are no RIC's. Thus when $|k|>2$ there are no RIC's for the standard map. Mather (Mather, 1984) refines this criteria using the explicit Lipschitz cone to obtain the bound $|k|> 4/3$. MacKay and Percival (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 (Stark, 1986) has shown that the criteria of MacKay and Percival are exhaustive: if there is no invariant circle the method will eventually show non-existence.

These bounds compare favorably with the numerical results of Greene (Greene, 1979) who presents numerical evidence that the last invariant circle of the standard map has a rotation number equal to the golden mean, and that it goes away at $k \approx 0.971635406$. In fact converse KAM theory is much better at producing practical results than the KAM theorem: as we mentioned before, Herman shows that the golden mean invariant circle of the standard map surely exists for $k<1/34$.

4 Variational Principles

In this chapter we show that the twist property guarantees the existence of a generating function, and a corresponding variational principle for the mapping. Conversely, any mapping which has a "Lagrangian variational principle" satisfies the twist property. The properties of the generating function will allow a straightforward proof of` the existence of periodic and quasiperiodic orbits and permit us to obtain many quantitative and qualitative properties of these orbits in the next few chapters.

4.1 Generating Function

Let T: $(x,y) \rightarrow (x',y')$ be the lift of a twist mapping to the plane $\mathcal{R}^2$. Then there exists a generating function $F(x,x')$ such that

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

or alternatively

$$dF(x,x') = y'dx' - ydx$$

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 $y(x,x')$. First lift the mapping to the plane $\mathcal{R}^2$, and consider the verticals $x = \xi$ and $x = \xi'$. The curve $T(x=\xi)$ intersects ξ' exactly once by the twist condition. Define $y'(\xi,\xi')$ to be this intersection. Similarly define $y(\xi,\xi')$ to be the unique intersection of $T^{-1}(x=\xi')$ with the vertical $x=\xi$.

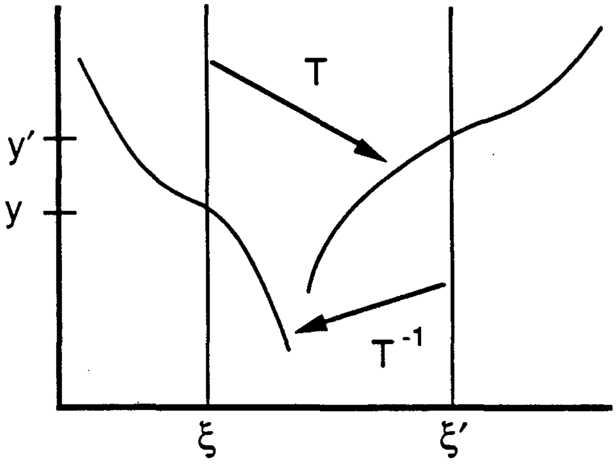


Figure 4.1.1 XBL 896-2122

For a given (x,x′) define the generating function by

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

where γ is a path which begins at some arbitrary point (x_0, x_0') and ends at (x,x′).

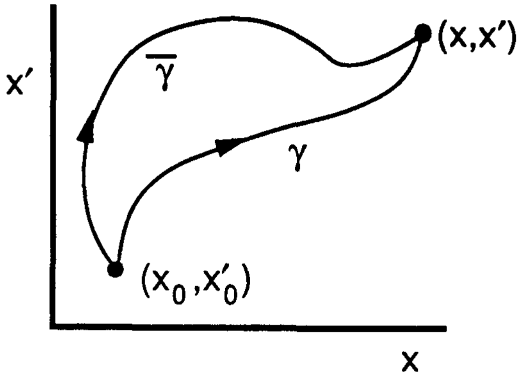


Figure 4.1.2 XBL 896-2123

This integral is independent of the choice of path: consider a second path $\overline{\gamma}$ which has the same endpoints as γ. By Stokes' theorem the integral $\int ydx$ around the closed loop $\gamma - \overline{\gamma}$ 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 $\int dy' \wedge dx' = \int y'dx'$ over this same loop. Thus the integrals of dF along γ and $\overline{\gamma}$ are equal, and dF is an exact one form.

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.

Furthermore the twist condition implies that

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

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

The mapping generated by F is area preserving because

$$dy \wedge dx = -F_{12}\, dx' \wedge dx$$
$$dy' \wedge dx' = F_{12}\, dx \wedge dx' = -F_{12}\, dx' \wedge dx$$

since $dy' = F_{12}dx + F_{22}dx'$ and $dy = -F_{11}dx - F_{12}dx'$.

This construction of the generating function yields a useful interpretation. Consider a curve $\mathcal{C}$ connecting z to $\bar{z}$, and its iterate $\mathcal{C}'$ connecting z′ to $\bar{z}'$.

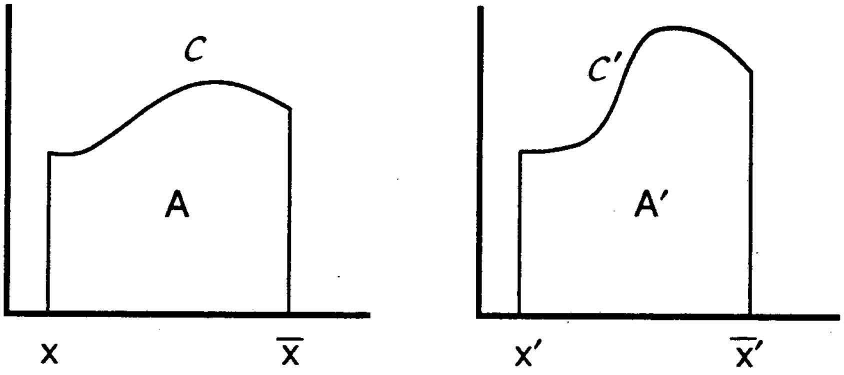


Figure 4.1.3 XBL 896-2124

The area under $\mathcal{C}$ is the integral $\int y dx$ along $\mathcal{C}$, while that under $\mathcal{C}'$ is $\int y' dx'$. Recalling that $dF = y'dx' - ydx$, we see that the difference between these areas is

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

We will find this relationship of great use in computing fluxes (see §7.7).

4.2 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 §3.2). A rotational circle is a curve which ranges from (x,y) to (x+1,y). Since the mapping is periodic $T\mathcal{C}$ ranges from (x′,y′) to (x′+1,y′). Thus the general formula for difference in areas becomes the net flux:

$$\mathcal{F} = F(x+1,x'+1) - F(x,x')$$

The net flux is zero if the generating function is a periodic function of $^1/_2(x+x')$; it can depend arbitrarily on x′-x. Such a mapping is called exactly symplectic because in this case the one form y′dx′ - ydx is exact on the cylinder: its integral is path independent even for paths that encircle the cylinder.

4.3 Examples

Standard Mapping: A generating function for the standard mapping is

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

This is the same as the energy function for the Frenkel-Kontorova model (see §1.3.3).

From another point of view the generating function is a discrete version of the Lagrangian for a dynamical system. For the standard map, this separates into the familiar form of kinetic minus potential energies, where the "velocity" is x′-x for the discrete time system, and the potential is $-k/4\pi^2 \cos(2\pi x)$. Thus we see that the standard map is a discrete approximation to the pendulum.

Note that the standard map has zero net flux.

Billiards: The generating function for a convex billiard is the function which gives the length between two boundary points. These can be labeled by the arc length coordinate, s; the generating function is

$$F(s,s') = \sqrt{(x-x')^2 + (y-y')^2}$$

$$\text{where} \quad \begin{array}{l} x = x(s),\ y = y(s) \\ x' = x(s'),\ y' = y(s') \end{array}$$

are two points on the boundary. The derivative of F with respect to s′ gives

$$\frac{\partial F}{\partial s'} = \frac{(x'-x)\frac{\partial x'}{\partial s'} + (y'-y)\frac{\partial y'}{\partial s'}}{\sqrt{(x-x')^2 + (y-y')^2}} = \cos(\theta')$$

This is the dot product of the unit vector pointing along the trajectory with the unit tangent vector to the boundary; that is, the cosine of the angle θ′ between them.

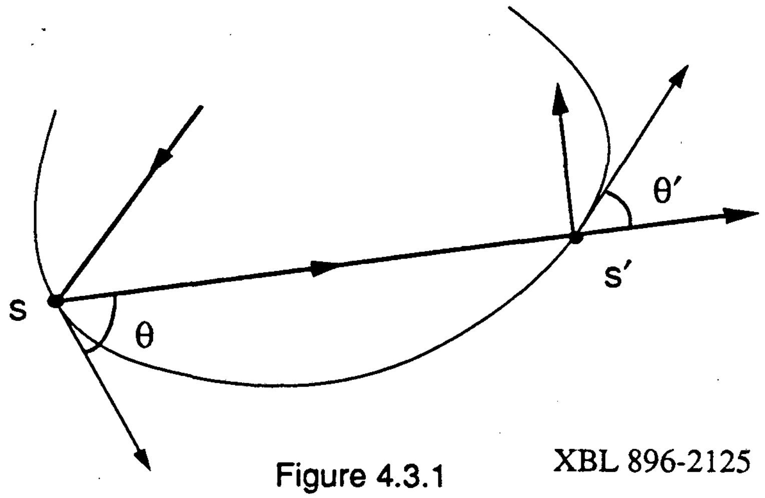


Figure 4.3.1 XBL 896-2125

Similarly F_1 is the negative of the cosine of θ. Thus the momentum coordinate is cosθ; this is natural since cosθ is the component of the particle momentum parallel to the boundary if we set the magnitude of p to one. In these coordinates the billiard map is area preserving.

The twist for the billiard is

$$\frac{\partial^2 F}{\partial s \partial s'} = \frac{\sin(\theta)\sin(\theta')}{F}$$

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 the these notes. To translate the Aubry-Mather theory for billiards replace minimizing by maximizing, and minimax by maximin. Note that the twist vanishes for glancing orbits, $\theta = 0$ or π.

The circle billiard has the generating function

$$F = 2r \sin\left(\frac{s'-s}{2r}\right)$$

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 the stadium, is left as an exercise to the reader!

4.4 Action

For a continuous time Lagrangian system, the action is the integral of the Lagrangian along a path q(t) in configuration space. Stationary points of the action with respect to variations in the space of paths q(t) with given endpoints $q(t_0) = x$, and

$q(t_1) = x'$ are orbit segments of this dynamical system. In fact this procedure can be used to construct the generating function for a mapping: suppose the Lagrangian depends periodically on time, with period $t_1 - t_0$, and $q(t)$ is an orbit segment for one period, then

$$F(x,x') = \underset{q(t)\ \text{stationary}}{\int_{t_0}^{t_1}} L(q,\dot{q},t)dt$$

where $F(x,x')$ depends only on the endpoints of the stationary segment. With this construction, it is clear that the action for the mapping is

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

and depends only on the configuration points corresponding to the intersections of the continuous system with a Poincaré section.

An orbit segment is a configuration $\{x_m, \ldots x_n\}$ which is a stationary point of the action, where x_m and x_n held fixed. Variation of the action gives the equations

$$\delta W = 0 \Rightarrow F_2(x_{j-1},x_j) + F_1(x_j,x_{j+1}) = 0$$

for $m< t<n$, which implies that the two definitions of momentum agree at each point on the orbit: $y'(x_{j-1},x_{j}) = y(x_j,x_{j+1}) = y_j$.

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

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

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 that $y_n=y_0$; therefore the orbit is indeed periodic.

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.

Examples: For the standard map, stationary points of the action must satisfy the equation

$$x_{t+1} - 2x_t + x_{t-1} = -\frac{k}{2\pi}\sin(2\pi x)$$

which is the Lagrangian form of the equations.

For the billiard, stationarity of the action implies that the angle of incidence equals the angle of reflection for each bounce.

5 Periodic Orbits

In this lecture and the next we will prove 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.

5.1 Minimizing orbits

The first variation of the action about an orbit is zero by definition. 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, however. Consider first a finite segment of a orbit, $\{x_m,...x_n\}$; let $\delta^2 W\{x_m,...x_n\}$ be the quadratic form obtained from the second order term in the expansion of W for fixed x_m and x_n:

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

If $\delta^2 W$ is non-negative for all non-zero vectors $\{\delta x_{m+1}, ...\delta x_{n-1}\}$, then the orbit segment is <u>locally minimizing</u>. If $\delta^2 W$ is positive definite, then the minimum is non-degenerate.

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

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

$$W\{\xi\} - W\{x\} \geq 0$$

If <u>every</u> <u>finite</u> segment of an orbit is minimizing then the orbit is minimizing. There are two reasons for allowing only deformations with compact support: firstly, if the deformation did not have compact support, the action difference $W\{\xi\} - W\{x\}$ would not necessarily be finite (being an infinite sum), and the two orbits could not be compared. Secondly, anchoring the asymptotic ($t \to \pm\infty$) behavior of the orbit acts as a kind of boundary condition, and we will find different minimizing orbits when different boundary conditions are imposed.

It is not obvious that minimizing orbits exist. We will first show that when ω is rational, the twist condition implies their existence. There are two steps: first we consider orbits which minimize $W_{(m,n)}$, and then we show that these also minimize W. In the next lecture we will consider irrational ω.

5.2 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 which 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 mapping 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

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

where B,C are positive.

Proof: Define the curve $\xi_\lambda = x + \lambda(x'-x)$, connecting x to x' as λ ranges from 0 to 1. On this curve

$$F(x,x') = F(x,x) + \int_0^1 d\lambda \; F_2(x,\xi_\lambda)\;(x'-x)$$

Repeating this construction on F_2 gives

$$F(x,x') = F(x,x) + \int_0^1 d\lambda\, F_2(\xi_\lambda,\xi_\lambda)\,(x'-x) - \int_0^1 d\lambda \int_0^\lambda d\mu\, F_{12}(\xi_\mu,\xi_\lambda)\,(x'-x)^2$$
$$\geq A - B|x'-x| + C|x'-x|^2$$

where $A = \min F(x,x)$ and $B = \max |F_2(x,x)| > 0$ exist by periodicity when the net flux is zero, and $C = {}^1/_{2K} > 0$, where K is the twist constant □

This result can be generalized to symplectic twist maps on $\mathcal{T}^N\times\mathcal{R}^N$. Here we assume that the mapping has uniformly positive definite twist: there exists a positive definite matrix C, such that $\delta x^t F_{12}(x,x')\,\delta x \leq -\,\delta x^t\, C\, \delta x$ for all x,x′. In this case the above proof of the growth condition can be directly transcribed (MacKay, Meiss and Stark, 1989).

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

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

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

In particular $W_{(m,n)} \geq n(A - {}^1/_4\, B^2/C)$; thus it has a lower bound.

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: each of the x_t for $0<t<n$ are bounded because the bound on $W_{(m,n)}$ implies that the above sum 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.

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 which minimizes $W_{(m,n)}$ □

The minimum is not unique. In fact by relabelling the origin of time, one minimum gives rise to n. In exceptional cases, such as the integrable twist mapping (e.g. the standard map with k=0) there is an entire curve of minima, forming the rational frequency invariant circle (I leave it as an exercise to explicitly demonstrate that these orbits do indeed minimize $W_{(m,n)}$).

5.3 Aubry's Fundamental Lemma

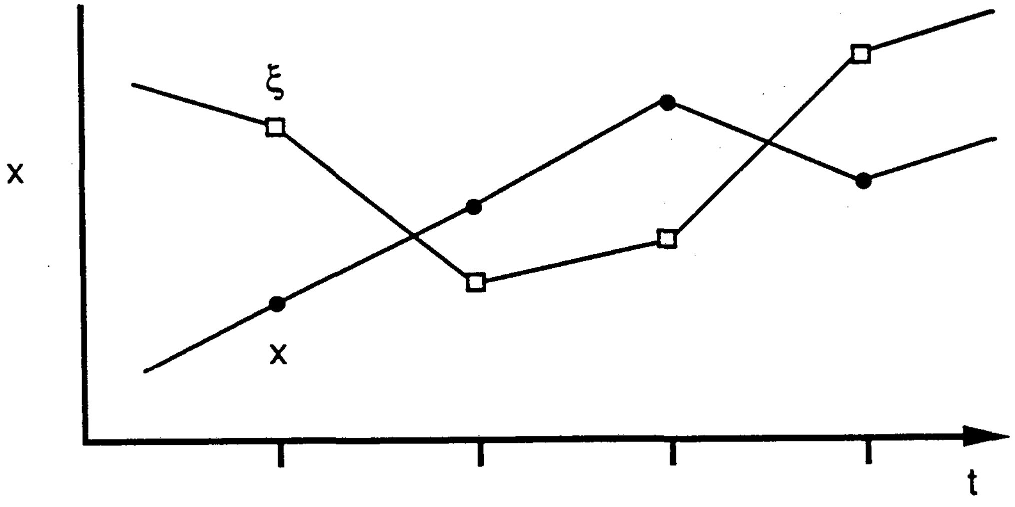
We have obtained periodic orbits which minimize the periodic action. In order to determine whether these orbits are truly minimizing orbits, we need a fundamental lemma of Aubry (Aubry, 1983).This lemma is the direct analogue of Morse's theorem (Morse, 1924) for geodesics (two minimum length geodesics for a Riemannian metric on the torus cross at most once) and a global version of the theorem that locally minimizing orbits have no conjugate points (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 space (x,t) and join successive points with straight lines to form the continuous curves

$$x(t) = (x_j - x_{j-1})(t-j) + x_j \qquad \text{for } j-1 \le t \le j$$

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.



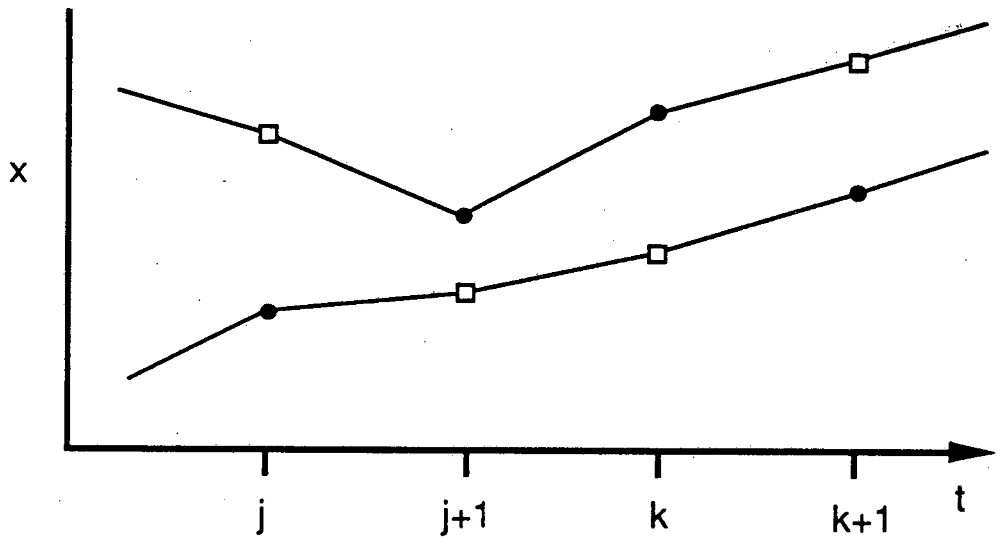
XBL 896-2128

Figure 5.3.1

Proof of Aubry's Fundamental Lemma: Suppose the converse, that {x} and {ξ} cross twice. We will obtain a contradiction. There are three possible cases: (i) the crossing points both occur at non-integer values of t, (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 {ξ} and show that at least one of these has smaller action, implying that both {x} and {ξ} cannot be minimizing. Let the two trajectories cross between times j and j+1 and times k and k+1. Define the deformations

$$\{\tilde{\xi}\} = \{\cdots\xi_{j-1},\ \xi_j,\ x_{j+1},\cdots x_k,\ \xi_{k+1},\ \xi_{k+2}\cdots\}$$
$$\{\tilde{x}\} = \{\cdots x_{j-1},\ x_j,\ \xi_{j+1},\cdots \xi_k,\ x_{k+1},\ x_{k+2}\cdots\}$$



XBL 896-2127

Figure 5.3.2

Note that it is necessary to have {x} and {ξ} cross twice to construct these deformations, because the definition of minimizing required that the variation occur only on a finite segment.

Consider the orbit segments running from time j to k+1. Since {x} and {ξ} were assumed to be minimizing the new segments must not have smaller action. Add the actions of these two segments to get

$$\begin{aligned} W\{\tilde{x}\} + W\{\tilde{\xi}\} = F(x_j, \xi_{j+1}) + W\{\xi_{j+1},\cdots\xi_k\} + F(\xi_k, x_{k+1}) \\ F(\xi_j, x_{j+1}) + W\{x_{j+1},\cdots x_k]\} + F(x_k, \xi_{k+1}) \end{aligned}$$

Subtract from this the sum W{x} +W{ξ}, grouping those terms referring to times j, j+1 and k,k+1 together:

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

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

$$F(x,\xi') + F(\xi,x') - F(x,x') - F(\xi,\xi') = \int_{\xi}^{x} d\lambda \int_{x'}^{\xi'} d\mu F_{12}(\lambda,\mu) < -C(x-\xi)(\xi' -x')$$

(In the case of no crossing, the inequality must be reversed because dλ dμ is effectively negative). Therefore after a crossing

$$F(x,\xi') + F(\xi,x') - F(x,x') - F(\xi,\xi') < 0$$

So difference between the actions of the modified orbits and the original orbits is

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

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

Case (ii) is proved similarly: the difference between the actions has contributions only from the non-integer 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 before. 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 $\tilde{\xi}_j = \xi_j$, $\tilde{\xi}_{j+1} = \xi_{j+1}$, but $\tilde{\xi}_{j+2} \neq \xi_{j+2}$, 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 ❑

<u>Corollary</u>: 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 they cross again ❑

In particular a minimizing periodic orbit can not cross any of its translates

$$x_t \rightarrow x_{t+k} + j$$

for any integers k,j because the relabelled orbit is also minimizing.

<u>Corollary</u>: Minimizing orbits are <u>monotone</u>: For all t, t′ and j, j′

$$x_t + j < x_{t'} + j' \;\Rightarrow\; x_{t+1} + j < x_{t'+1} + j'$$

Proof: Let $x_t + j \rightarrow x_t$ and $x_{t'} + j' \rightarrow \xi_t$ and apply the fundamental lemma: ξ_t cannot cross x_t ❑

5.4 Minimizing (m,n) orbits

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 which minimizes $W_{(m,n)}$. We must show that there is no infinite configuration which has smaller action. For example consider an orbit {ξ} of type (km,kn) which minimizes $W_{(km,kn)}$. By the fundamental lemma, this orbit cannot cross any of its translates. Now suppose {ξ} 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 $\xi_{t+n} - m > \xi_t$ or $\xi_{t+n} - m < \xi_t$. 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 shown that if $\{x_0,x_1,...,x_{n-1}\}$ minimizes $W_{(m,n)}$ then $\{x_0,x_1,....x_{kn-1}\}$ = $\{x_0,x_1,...,x_{n-1},x_0+m,x_1+m, ...,x_{n-1}+(k-1)m\}$ minimizes $W_{(km,kn)}$ for all k>1. Since the segment $\{x_0,x_1,....x_{kn-1}\}$ is minimal, its action must be less than that of any variation $\{x_0, x_1+\delta x_1,..., x_{kn-1} + \delta x_{kn-1}, x_{kn-1}\}$ with the same endpoints. 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 is not true in higher dimensions. Hedlund gives a counter-example for the case of geodesics on a three-torus (Hedlund, 1932). 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 locally minimizing orbits may still play an important role (Kook and Meiss, 1988).

5.5 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 must be other critical points.

To show this more formally we construct a new orbit by constrained minimization. Recall that 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. Thus a minimizing orbit has index zero.

Theorem: For every (m,n) there exists a second periodic orbit. Typically this orbit has index one. This is the minimax orbit.

Proof: Any translate $\{\xi_t\}$ of $\{x_t\}$ does not cross $\{x_t\}$ since both orbits are minimizing. Choose the translate ξ_0 which is closest to x_0. Now choose a path $\zeta(\lambda) = \{x_0(\lambda),....x_{n-1}(\lambda)\}$ for $\lambda \in [0,1]$ connecting these two neighboring minima of $W_{(m,n)}$. Since $W_{(m,n)}$ is continuous it must have a maximum along this path. In Figure 5.5.1 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 shaded point.

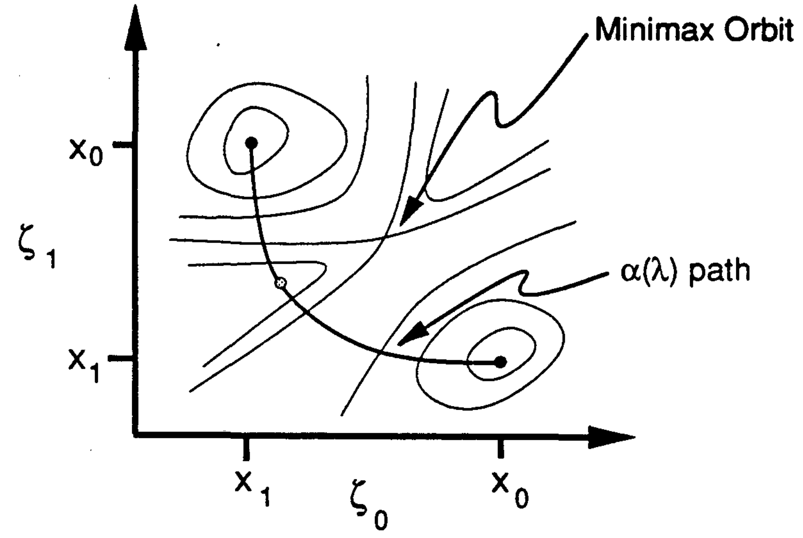


Figure 5.5.1 XBL 896-2126

Finally find the smallest of such maxima by varying the path $\zeta(\lambda)$. This gives an (m,n) periodic orbit because at the maximum $\delta W_{(m,n)} = 0$. This minimum over the maxima is the minimax orbit □

The minimax orbit is well ordered with respect to the minimizing orbit. These two form the "island chain" structure so commonly seen in pictures of area preserving maps. In fact one can see that the residue of a non-degenerate minimizing orbit must be negative (MacKay and Meiss, 1983), indicating that it is hyperbolic (§2.2). On the other hand, the residue of a non-degenerate minimax orbit must be positive, so that it is either elliptic or hyperbolic with reflection. The first case gives the familiar island chain structure. The latter occurs after the elliptic orbit undergoes a period doubling bifurcation (Greene et al., 1981), and signals 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 (MacKay, Meiss and Percival, 1987).

This completes the analysis of the ordered periodic orbits. There are of course may other periodic orbits in a typical mapping. Some of these can be understood by similar techniques to the above, for example those which are librating within an island chain. These 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 (§1.3.2), the above theorems prove the existence of librating periodic orbits for all rational frequencies in some interval. Thus we obtain both minimizing and minimax trapped periodic orbits; we call these orbits of class one (rotational orbits have class zero). If these minimax periodic orbits are elliptic, then librating orbits of class two, and so forth can be obtained. This leads to the picture of islands around islands, *ad infinitum* (Meiss, 1986).

6 Quasiperiodic Orbits

We will 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 an irrational, $m/n \to \omega$. This approach was pioneered in the numerical experiments of John Greene, who determined the break-up of invariant circles by studying the stability of nearby periodic orbits (Greene, 1968; Greene, 1979).

Recall from §2.4 that a quasiperiodic orbit is one with irrational rotation number which is recurrent. An orbit $\{x_t\}$ is recurrent if every point on the orbit, x_t, can be obtained as a limit of other points on the orbit, i.e. there exists a sequence of times t_1, t_2, ... $\to\infty$, such that $x_{t_k} \to x_t$ as $k \to \infty$. We will find both quasiperiodic orbits and orbits which are not recurrent, but which are homoclinic to quasiperiodic orbits.

As we will see the quasiperiodic orbits either densely cover a rotational invariant circle or an invariant Cantor set, a "cantorus." A cantorus is the remnant of an invariant circle.

6.1 Invariant Circles are Minimizing

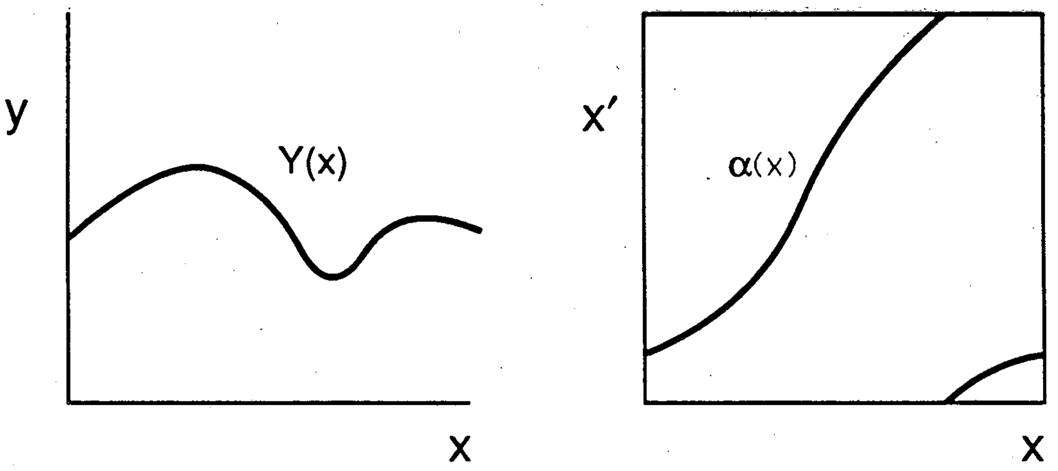
In a discussion of rotational invariant circles of twist maps, the concept of minimizing orbits arises naturally since:

Theorem (Mather): Every orbit on a rotational invariant circle is minimizing.

Proof: By Birkhoff's theorem, every RIC is the graph of a Lipschitz function Y(x). Let $S(x) = \int Y(x)\,dx$, integrating from some arbitrary point. Since Y(x) is periodic, S(x) is as well. Define the function

$$H(x,x') = F(x,x') - S(x') + S(x)$$

Let the orbit on the invariant circle be described by $x' = \alpha(x) = T_x(x,Y(x))$.



XBL 896-2141

Figure 6.1.1

By the twist condition (see p. 25) we have

$$H_1 = F_1(x,x') + Y(x) \quad = 0 \text{ iff } x' = \alpha(x)$$
$$H_2 = F_2(x,x') - Y(x') \quad = 0 \text{ iff } x' = \alpha(x)$$

which implies that $H(x,\alpha(x)) = H_0$ is constant, and that all critical points of H occur on $x' = \alpha(x)$. Furthermore, $H(x,x')$ satisfies the same growth condition (see p. 32) as $F(x,x')$, and zero net flux implies that $H(x+1,x'+1) = H(x,x')$; so that H is bounded from below. Therefore

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

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

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

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. In fact the invariant torus must be a Lagrangian manifold as well (Herman, 1988).

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

Proof: The fact that the set we are considering is an invariant Lagrangian graph implies that $Y(x) = \nabla S(x)$. Use this S to define H as before, and the generalization of the growth condition to follow the above proof ❑

6.2 Monotone Sets

We can generalize the definition of monotone from an orbit, as in §5.3, to any invariant set. Let π be the projection operator onto the x axis: $\pi(z) = x$, then an invariant set M is monotone if for all $z, \zeta \in M$,

$$\pi(z) < \pi(\zeta) \Rightarrow \pi(T(z)) < \pi(T(\zeta))$$

An orbit is monotone if the set formed from all its translates is monotone: for all m,m′ and t, t′

$$x_t + m < x_{t'} + m' \Rightarrow x_{t+1} + m < x_{t'+1} + m'$$

We showed in §5.3 that Aubry's fundamental lemma implies that minimizing periodic orbits are monotone. Monotone sets for twist maps have nice properties:

Lemma: A monotone set is a graph over x.

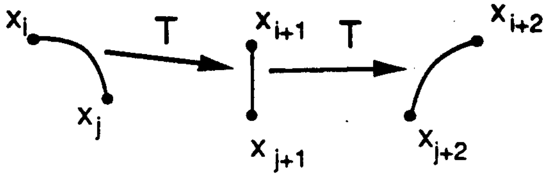
Proof: Suppose that there are two points, $z= (x,y)$ and $\zeta = (\xi,\eta)$ in a monotone set which have the same x value: $x= \pi(z) = \xi = \pi(\zeta)$, but different y values, say $y > \eta$. The twist condition implies that $\pi(T^{-1}z) < \pi(T^{-1}\zeta)$. This violates the definition of monotone. Thus if $x=\xi$, then $y = \eta$, and $\pi(z) = \pi(\zeta) \Rightarrow \pi(T(z)) = \pi(T(\zeta))$ ❑

Lemma: Any limit of monotone orbits is monotone

Proof: suppose that for each k $\{x^{(k)}\}$, is a monotone orbit. Then, for each k, $x^{(k)}_i < x^{(k)}_j \Rightarrow x^{(k)}_{i+1} < x^{(k)}_{j+1}$, and in the limit,

$$x^{(\infty)}_i < x^{(\infty)}_j \Rightarrow x^{(\infty)}_{i+1} \leq x^{(\infty)}_{j+1}$$

Suppose equality occurs above, then the twist condition implies that $x^{(\infty)}_{i+2} > x^{(\infty)}_{j+2}$,



XBL 896-2155

Figure 6.2.1

but this implies there is a K such that for all k > K, $x^{(k)}_{i+2} > x^{(k)}_{j+2}$, contradicting the assumption that $\{x^{(k)}\}$ is monotone for all k. Thus equality can not occur and the limit must be monotone ❑

Lemma: The closure of a monotone 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

$$x_0 < \xi_0 \Rightarrow x_n \leq \xi_n$$

However, as above, 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. First we show that the restriction of the twist mapping to a monotone set is equivalent to a mapping on the circle:

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 $x \in \mathcal{R}$ satisfying $\alpha(x+1) = \alpha(x)+1$.

Proof: The closure of M is monotone, and 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 $\mathcal{R}$ by periodicity. Thus α is continuous, and because T is invertable, it has a continuous inverse ❑

Below is shown the construction of $\alpha(x)$ for a (2,5) minimizing orbit. For an (m,n) orbit there are m inequivalent translations in the (x,y) plane. For the (2,5) case they are $\{x\} = \{...,x_0, x_1, x_2, x_3, x_4,...\}$, and $\{\xi\} = \{...,x_3-1, x_4-1, x_0+1, x_1+1, x_2+1,...\}$. We show part of the real line (of length 2) below, and the 5 points of each of the orbits $\{x\}$ and $\{\xi\}$ which lie in this segment. Define the function $\alpha(x)$ on the orbit points such that x_t is mapped to x_{t+1}, and similarly to take ξ_t to ξ_{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.

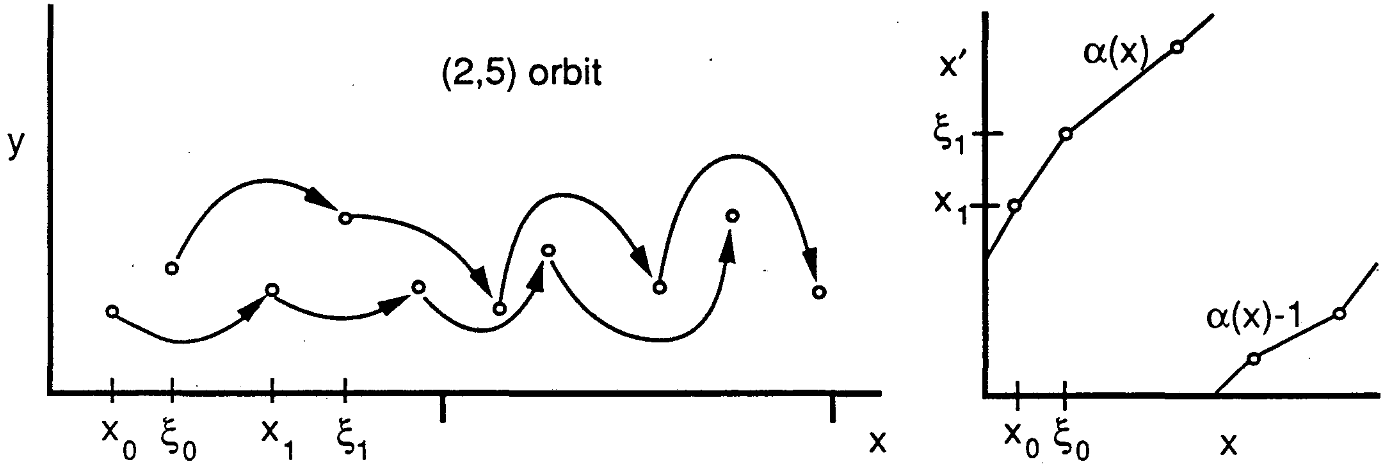


Figure 6.2.2 XBL 896-2140

To see how this fails for a non-monotone orbit, consider a configuration of type (2,4). Recall from §5.4 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). Below we sketch such an orbit. Note that 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. We see according to the picture that although $\xi_1 > x_1$, $\xi_2 < x_2$. This is reflected by a non-monotonic segment in the induced $\alpha(x)$, which is therefore not a homeomorphism.

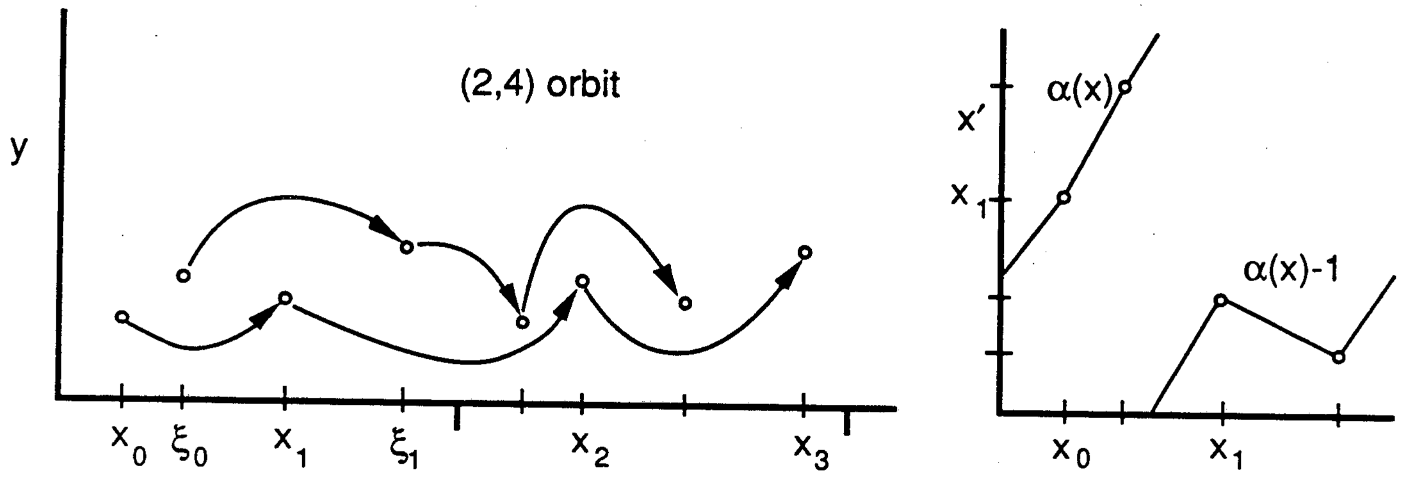


Figure 6.2.3

XBL 896-2143

A classic theorem of Poincaré implies that every orbit of a homeomorphism of the circle has a rotation number, and the rotation number is the same for all orbits–see Theorem (B2) in Appendix B. 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 is an M and δ such that $|x_t - \xi_t| < \delta$ for all $0 \leq t \leq M$. From lemma (B1) in Appendix B it follows that $|x_t - x_0 - t\omega| < 1$, and similarly for ξ. Thus

$$
\begin{gathered}
|x^t - \xi^t - (x-\xi) - t[\omega(x)-\omega(\xi)]| \leq 2 \\
\Downarrow \\
|x^t - \xi^t| \geq t|\omega(x)-\omega(\xi)| - |x-\xi| - 2 \\
\Downarrow \\
|\omega(x)-\omega(\xi)| \leq 2(1+\delta)/M
\end{gathered}
$$

Now consider a sequence of monotone states $\{x^{(k)}\}$, with periods $n^{(k)} \to \infty$, such that $m^{(k)}/n^{(k)} \to \omega$. Assume that $\{x^{(k)}\}$ approach a limit $\{x^{(\infty)}\}$, so that there is a K and δ for which $|x^{(k)}{}_t - x^{(\infty)}{}_t| < \delta$ for all $0 \leq t \leq n^{(k)}$ and $k \geq K$. Since the periods go to infinity, the above inequality implies that the rotation number of $\{x^{(\infty)}\}$ is the same as the limit of the rotation numbers of the $\{x^{(k)}\}$ ❑

6.3 Existence of Quasiperiodic Orbits

In §5.4 we proved that there is a minimizing monotone state for every (m,n). We now show that this is true for all ω.

Theorem (Aubry and Le Daeron, 1983; Mather, 1982) 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 preceding lemmas 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 $\xi^{(k)}{}_i = x^{(k)}{}_i$ and $\xi^{(k)}{}_j = x^{(k)}{}_j$, $\{\xi^{(k)}\} \to \{\xi\}$. Let

$$\varepsilon^{(k)} = \max(|x^{(k)}{}_t - x_t|, |\xi^{(k)}{}_t - \xi_t|) \quad \text{for} \quad i \leq t \leq j$$

Since F(x,x′) is differentiable, the difference between the action of the segments $\{x^{(k)}\}$ and $\{x\}$ is bounded by a multiple of $\varepsilon^{(k)}$:

$$\left|W\{x_i^{(k)}, \cdots x_j^{(k)}\} - W\{x_i, \cdots x_j\}\right| \leq K(i-j)\varepsilon^{(k)} \qquad \forall k > N$$

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

$$W\{\xi_i, \cdots \xi_j\} - W\{x_i, \cdots x_j\} \geq W\{\xi_i^{(k)}, \cdots \xi_j^{(k)}\} - W\{x_i^{(k)}, \cdots x_j^{(k)}\} - 2K(i-j)\varepsilon^{(k)} \qquad \forall k > N$$

Now because $\varepsilon^{(k)} \to 0$ as $k \to \infty$, and each $\{x^{(k)}\}$ is minimizing, we have

$$W\{\xi_i, \cdots \xi_j\} - W\{x_i, \cdots x_j\} \geq 0$$

and so the limit is minimizing ❑

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. We will show below that the totality of minimizing states for a frequency ω is a closed monotone set.

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 which are not recurrent. We will discuss these below when we consider heteroclinic orbits.

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 John Greene in his studies of the break-up of invariant circles (Greene, 1979). He found, for example that when the residues of periodic orbits with $m_i/n_i \rightarrow \omega$ approach zero, an invariant circle exists, while when the residues are unbounded, there is no invariant circle. An optimum way to take this limit is using the continued fraction convergents of ω. This leads to rapid convergence, especially in the unstable case. We observe that the residue of the convergents approaches infinity as the Lyapunov multiplier of the quasiperiodic orbit to the power of the period: $|R| \sim (\lambda/4)^n$. Properties of the quasiperiodic orbit, such as the position of particular points, are approached as the inverse of the residue: $x^{(i)} - x^{(\infty)} \sim 1/\lambda^n$; thus this approach is extremely rapid,and for moderately unstable orbits, good approximations are obtained for low periods (MacKay, Meiss and Percival, 1984).

6.4 Cantori

We have seen that quasiperiodic minimizing orbits exist for all ω, for any twist mapping. The reader should not be surprised about this result in the case when 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 (§2.6). 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 parameters are sufficiently strong (§3.3.4). So the question is: 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 1) an invariant circle or 2) an invariant Cantor set.

Proof: Since the minimizing orbit is monotone, its closure can be extended to a homeomorphism, $\alpha(x)$ of the circle. It is a standard result that when such homeomorphisms have irrational rotation number, the set of limit points of the orbit of any point is unique, invariant, and is either is the entire circle or a Cantor set [see appendix B, (B3)]. 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 for twist maps "cantori". He suggested the existence of cantori based on a variational principle for quasiperiodic orbits (Percival, 1979); Aubry independently suggested their existence (Aubry, 1978). Furthermore, Percival showed that cantori do exist for a particular family of maps: the sawtooth maps (Aubry, 1983; Percival, 1979).

A cantorus is an invariant set which is "trying" to be a rotational invariant circle, however orbits on this set fail to cover the circle: they omit a countable set of open intervals, or gaps. If one imagines the mapping as arising from the Poincaré section of a flow, the cantorus is a two dimensional torus with at least one deleted ribbon which winds around the torus with irrational helicity; hence the name.

Because orbits on the cantorus are ordered, the endpoints of a gap do not cross upon iteration: we can define a left endpoint x^l_0 and right endpoint x^r_0 to each gap such that $x^l_0<x^r_0$, and $x^l_t <x^r_t$. Each iterate of a gap gives a distinct gap because the orbit is quasiperiodic. The total length of the gaps (in x) is at most one, so the length of the iterate of any gap must eventually go to zero. Thus the orbit of any point which stays in the gaps for $t>t_0$, is asymptotic to the cantorus as $t\to\infty$.

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 there is 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, 1988)). They can then be labelled by t. There is a largest gap, we call this the principal gap, and use t=0 to designate it.

Cantori are typically hyperbolic, though I do not know of any theorem which guarantees this (when k is large enough all cantori of the standard map are hyperbolic (Goroff, 1985)). 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 smoothly from one when a cantorus is formed.

When the Lyapunov multiplier is larger than one, one can see that the iterate of any gap has a length which eventually must approach zero as λ^{-n}. This implies further that the Hausdorff dimension of the cantorus is zero (MacKay, 1988). This is remarkable, since it implies that when an invariant circle breaks its length falls immediately to zero; furthermore, its dimension discontinuously changes from one to zero (providing it becomes hyperbolic).

6.5 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 which are not covered by these results. Here we sketch the classification of the complete set of minimizing orbits.

Every minimizing orbit does have a frequency ω, and for each ω the set of minimizing orbits is monotone. In addition to the periodic and quasiperiodic minimizing orbits, a new class, the non-recurrent orbits must be considered.

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.

Aubry's fundamental lemma implies that periodic minimizing orbits are monotone. This can be generalized to any minimizing orbit:

Theorem: If $\{x\}$ is a minimizing orbit, then $\{x\}$ is monotone. Thus $\{x\}$ has a frequency ω.

Proof: see (MacKay and Stark, 1985)

Theorem: Let M_ω be the set of minimizing orbits with irrational frequency ω. Then M_ω is monotone. If $\{x\}$ and $\{\xi\}$ are quasiperiodic minimizing orbits they have the same closures: $M_\omega\{x\} = M_\omega\{\xi\}$. If $\{x\}$ is a non-recurrent minimizing orbit with frequency ω, then it lies in the gaps of $M_\omega\{x\}$.

Proof: see (MacKay and Stark, 1985)

Corollary: If there exists a rotational invariant circle with irrational frequency ω, then every minimizing orbit of frequency ω is recurrent, and densely covers the circle.

Proof: see (MacKay and Stark, 1985)

6.6 Heteroclinic Orbits

There are non-recurrent minimizing orbits for each ω. By Aubry's fundamental lemma, these must fall in the gaps of the recurrent orbits, and these orbits are heteroclinic to the gap endpoints. When ω is rational these orbits are crossing points of the stable and unstable manifolds of the minimizing orbit, the heteroclinic tangle formed from the separatrix. When ω is irrational these orbits fall in the gaps of the cantorus.

6.7 Mather's DW

The non-existence of an invariant circle is implied by the existence of a non-minimizing 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 $\{\zeta^{(k)}\}$ 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

$$\Delta W_\omega = \lim_{k\to\infty} \left[W_{(m_k,n_k)}\{\zeta^{(k)}\} - W_{(m_k,n_k)}\{x^{(k)}\} \right]$$

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

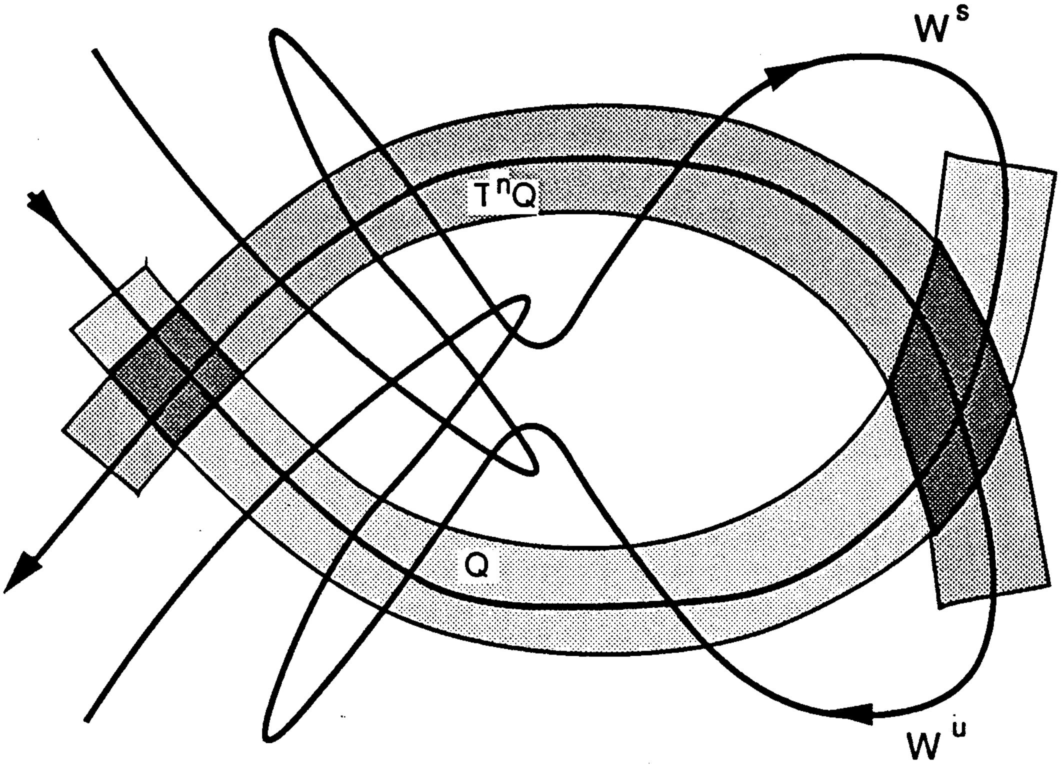
We will see in §7.11 that the quantity ΔW can be interpreted as the flux through the minimizing set. It is therefore natural that $\Delta W=0$ when there is an invariant circle.

7 Flux

7.1 Irregular Components

An irregular component, as defined by Birkhoff, is the complement of the elliptic periodic orbits, invariant circles, and periodic invariant circles (that is an invariant set which consists of a family of circles which map onto each other, as in an island chain). By Birkhoff's theorem (§3.3.3) any irregular component which contains a rotational circle, e.g. an irregular annulus, is bounded by rotational invariant circles. Furthermore Birkhoff's theorem implies that there is an orbit which goes from arbitrarily close to the lower RIC to arbitrarily close to the upper RIC.

If the stable and unstable manifolds of a periodic orbit have a transversal intersection, then there is an irregular component. A horseshoe of irregular orbits can be constructed by following the lobes of the stable and unstable manifolds. Existence of a horseshoe implies that there is a zero measure set of orbits which act chaotically: they can be mapped onto a Bernoulli shift. There are no results which imply that a non-zero measure of the orbits is chaotic in some sense, and it is not known whether irregular components typically have non-zero measure. There are examples of completely ergodic systems, for example the Arnol'd cat map (which is a twist map) and specially constructed examples of systems with both invariant circles and irregular regions (Wojtkowski, 1981).



XBL 896-2142

Figure 7.1

On the other hand, numerical evidence implies that the measure of a typical irregular component is non-zero; they seem to be “fat fractals” (Umberger and Farmer, 1985). A prime goal of a theory of transport is to understand how orbits move from one portion of an irregular component to another.

7.2 Flux–Definition

Our primary technique for studying transport is the flux of trajectories across surfaces in phase space (MacKay, Meiss and Perciva, 1984). To define the flux consider a closed loop $\mathcal{C}$, and its iterate $T\mathcal{C}$.

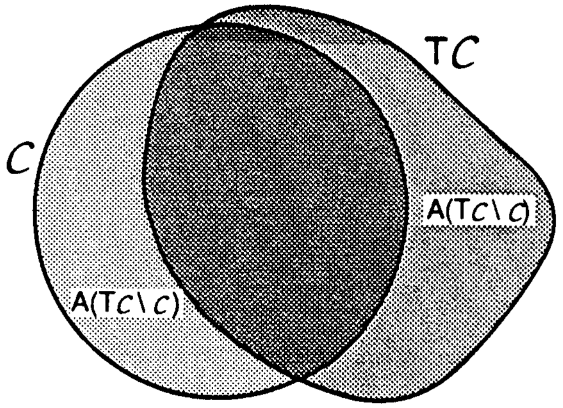

T$\mathcal{C}$
$\mathcal{C}$
A(T$\mathcal{C}$\ $\mathcal{C}$)
A(T$\mathcal{C}$\ $\mathcal{C}$)

Figure 7.2.1 XBL 896-2129

The flux through $\mathcal{C}$ is defined as the area inside T$\mathcal{C}$ which is also outside of $\mathcal{C}$:

$$\mathcal{F}(\mathcal{C}) = A(T\mathcal{C} \setminus \mathcal{C})$$

If $\mathcal{C}$ encloses finite area, then area preservation implies that this outward flux is the same as the inward flux, A($\mathcal{C}$\T$\mathcal{C}$). Similarly if the loop $\mathcal{C}$ is a rotational circle, the upward flux is the area which is above $\mathcal{C}$ and below T$\mathcal{C}$. If the map has zero net flux (§3.2), the upward and downward fluxes are equal.

If $\mathcal{C}$ is an invariant circle, then it has zero flux (recall §3.1).

Every 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 the loop $\mathcal{C}$: If motion in $\mathcal{C}$ is "random" in some sense then a trajectory will be trapped within $\mathcal{C}$ for a typical time

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

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 is replaced by the area of the connected irregular component inside $\mathcal{C}$. However, this is difficult to determine.

7.3 The Important Question

To best estimate transport rates through phase space, we would like to find which curves $\mathcal{C}$ have the minimum flux. For example 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 is infinite.

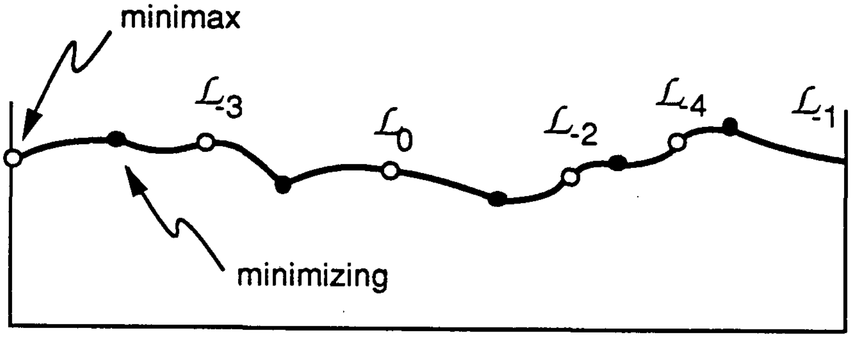
More generally, the transit time is most influenced by rotational circles which have small flux, and so we ask the question

What rotational circles have minimum flux?

Imagine varying the circle $\mathcal{C}$ over the annulus and minimizing the flux. The answer to this question will be discussed below. First lets discuss techniques for computing flux.

7.4 Partial Barriers and Turnstiles

A curve which in some sense approximates a rotational invariant circle acts as a partial barrier for the dynamics. One way to construct such a curve is to use the minimizing and minimax periodic (m,n) orbits (MacKay, Meiss and Perciva, 1984). The reason for this choice is that in the integrable limit these orbits do lie on a rotational invariant circle, and so one might expect them to approximate such a circle in general. Furthermore the monotonicity of the orbits allows a simple construction of a rotational (non-invariant) circle: choose a gap in the minimizing orbit, lets call it the principal gap, and fill it with an arbitrary curve, $\mathcal{L}_0$, which also goes through the minimax orbit. Take n-1 preimages of this curve to fill in the remaining gaps. The resulting curve is a rotational circle, and we call it a "partial barrier." Below is a partial barrier for a (2,5) orbit.



XBL 896-2130

Figure 7.4.1

This curve defines a boundary and divides the cylinder into two halves. To move from one to the other, a trajectory must cross the partial barrier. It can do so because the partial barrier is not an invariant curve (except in special cases): when it is iterated once each of the $\mathcal{L}_t$ becomes $\mathcal{L}_{t+1}$, except for $\mathcal{L}_0$. To visualize the flux through the partial barrier, take one more preimage of $\mathcal{L}_{-n+1}$, to obtain a second curve in the gap "0". It must connect the gap endpoints, because the endpoints lie on a periodic orbit.

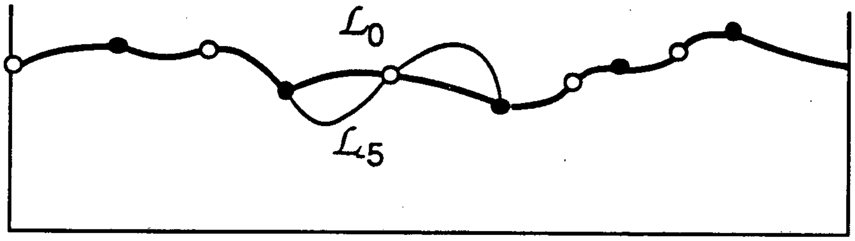


Figure 7.4.2 XBL 896-2131

The region below $\mathcal{L}_0$ and above $\mathcal{L}_{-n}$ is the region which crosses the partial barrier from below to above on one iteration of the mapping. Similarly the region below $\mathcal{L}_{-n}$ but above $\mathcal{L}_0$ is crosses from above the partial barrier to below upon one iteration.

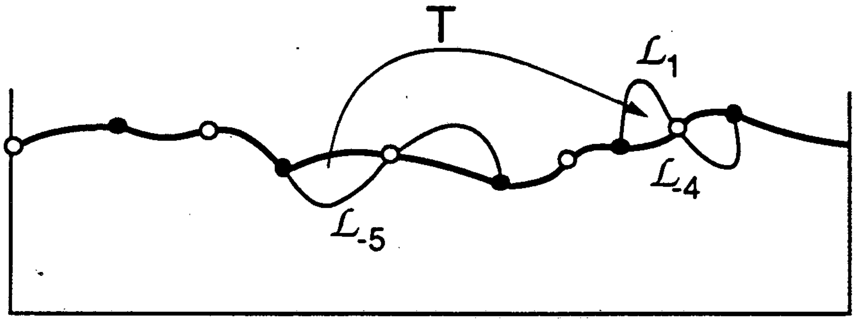


Figure 7.4.3 XBL 896-2134

The only way a trajectory can cross from below the partial barrier to above is to land in the lobe below $\mathcal{L}_0$ and above $\mathcal{L}_{-n}$. Because the net flux is zero, the flux up and down are equal, therefore $\mathcal{L}_{-n}$ and $\mathcal{L}_0$ must cross at least once, giving the characteristic figure-of-eight structure, which we call a turnstile (MacKay, Meiss and Perciva, 1984). This is because it acts as a "rotating door", dumping all the area in its left lobe above the partial barrier, and all the area in its right lobe below the partial barrier each iteration.

We would like to obtain the partial barrier through the (m,n) orbits which has the minimum flux, because this will give the best estimate of transport rates. However, as we will see below the flux is independent of the choice of the curve L_0, providing it connects two neighboring points on the minimizing orbit, and goes through the minimax orbit.

Turnstiles can be more complicated than we indicated above. For example there is nothing which prevents the turnstile from looking like:

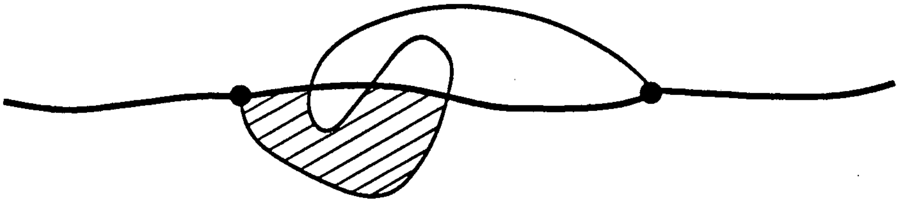


Figure 7.4.4 XBL 896-2133

In this case the flux is the striped region shown. Though I've never seen a turnstile with this structure, they may well occur in physical examples.

7.5 Heteroclinic Orbits

Probably the most familiar form of turnstile occurs in the construction of the stable and unstable manifolds of a hyperbolic orbit. Consider, for example the rotational (m,n) minimizing orbit. This orbit has two minimizing homoclinic orbits, the right-going and left-going. Consider the right-going one. A partial barrier is formed by choosing a gap (the principal gap) in this orbit, labelled with t=0, and connecting it with a segment of the unstable manifold, $\mathcal{U}_0$. This segment naturally goes through the minimax homoclinic orbit. Preimages of this gap converge to the minimizing (m,n) orbit, and are connected with unstable manifold, $\mathcal{U}_t$, t < 0. Images are connected with the stable manifold, $\mathcal{S}_t$, t > 0. In this way we obtain a piecewise smooth rotational partial barrier, with a discontinuity in slope at the right endpoint of the initial gap.

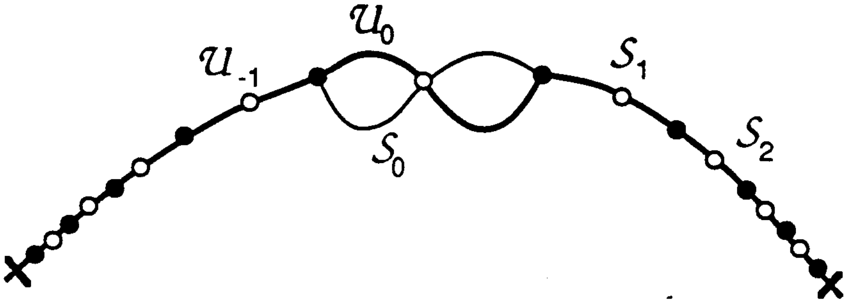


Figure 7.5.1 XBL 896-2132

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. The area between $\mathcal{S}_0$ and $\mathcal{U}_0$ crosses the partial barrier each iteration.

7.6 Cantori

A similar partial barrier can be obtained for the case of a cantorus. Choose a gap, and label it "0". 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. We can construct a partial barrier from two segments: one of which is asymptotic to the gap endpoints as $t \to -\infty$ ($\mathcal{U}_0$) and the other as $t \to \infty$ ($\mathcal{S}_0$). Set $\mathcal{L}_t = T^t(\mathcal{S}_0)$ for $t > 0$ and $\mathcal{L}_t = T^t(\mathcal{U}_0)$ for $t \leq 0$. These curves necessarily go through the minimax orbit, since it is homoclinic to the cantorus. If there is only one family of gaps, then the resulting curve is a rotational circle, and forms 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 the gap "0", but does not necessarily coincide with $\mathcal{U}_0$.

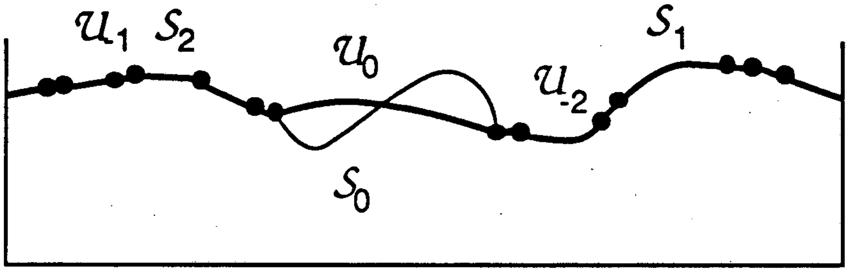


Figure 7.6.1

XBL 896-2136

Indeed if it did, then there would be an invariant circle, contrary to assumption. So the combination of $\mathcal{S}_0$ and $\mathcal{U}_0$ form a turnstile. They must cross at least once on the minimax orbit, and may do so several times. The flux through the cantorus is defined as before.

7.7 Resonances

For a resonance, an upper partial barrier can be constructed from the collection of all the past unstable and future stable segments connecting M^+_t to M^+_{t+n}, where M^+_t is the right-going, minimizing homoclinic orbit. Similarly the lower partial barrier is constructed from the left-going minimizing homoclinic orbit, M^-_t.

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 which we have constructed, depends on the choice of "t" at which we switch from unstable to stable segments. However, it is easy to see that the turnstile areas, and the total area of the resonance itself is independent of this choice.

7.8 Areas and Actions–the 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 use numerical integration; however, this is not the best way. Below I will show how these can be obtained from the action principle for the orbits making up the partial barriers (MacKay, Meiss and Percival, 1987).

The basic formula relating action to area was discussed in §4.1. Here we recall this result. Let $\mathcal{C}$ be a directed curve on the phase plane. Parameterize it by λ ranging over [0,1] , so that

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

Let A be the <u>algebraic area</u> "under" $\mathcal{C}$, i.e. the signed area bounded by the loop formed from $\mathcal{C}$, the vertical lines x = x(0) and x = x(1) and the horizontal line y=0 . The direction of the loop is set as that of increasing λ. Regions encircled by (counter) clockwise loops are defined to have (negative) positive area. For the simple situation depicted in §4.1, A is merely the <u>geometric area</u>. If, however, $\mathcal{C}$ intersects itself or if y(λ) 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 under $\mathcal{C}$. In any case we will still refer to A as the area "under" $\mathcal{C}$, though some regions may be included with negative sign.

The image of $\mathcal{C}$ under the map T is $T(\mathcal{C}) = \mathcal{C}'$. The area under $\mathcal{C}'$ is denoted A′, and, as is shown in §4.1, is the signed area enclosed by $\mathcal{C}'$, the verticals x=x′(0) and x=x′(1), and the horizontal y=0.

Let F(x,x′) be the generating function of the twist map T from the initial point with angle x(λ) to its image point with angle x′(λ). By definition

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

Integrating both sides with respect to λ, we obtain

$$\Delta F = F[x(1),x'(1)] - F[x(0),x'(0)]$$
$$= A' - A$$

This is the basic formula from which all the others follow.†

7.9 Flux–Periodic Orbits

The flux through the turnstile in a pair if (m,n) periodic orbits is easily obtained from the fundamental formula. Let A_0 be the area under the segment $\mathcal{L}_0$ connecting the point M_0 on the minimizing orbit to S_0 on the minimax orbit (in this chapter we denote phase space points with capital letters, and the corresponding configuration with lower case letters). Similarly A_t represents the iterates of this area. The fundamental formula implies that

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

The area of the turnstile is given by $A_0 - A_{-n}$, which is obtained by iteration of the above formula:

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

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

7.10 Stable and unstable segments

To derive formulae for the flux through homoclinic orbits, and cantori, it is convenient to consider a more general case. Two points X_0, and Z_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:

† Bob Easton has shown me that this relation, and the others which follow can be generalized to maps which do not satisfy the twist condition, and even those which are not area preserving.

$$\lim_{t\to\infty} |X_t - Z_t| = 0$$

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

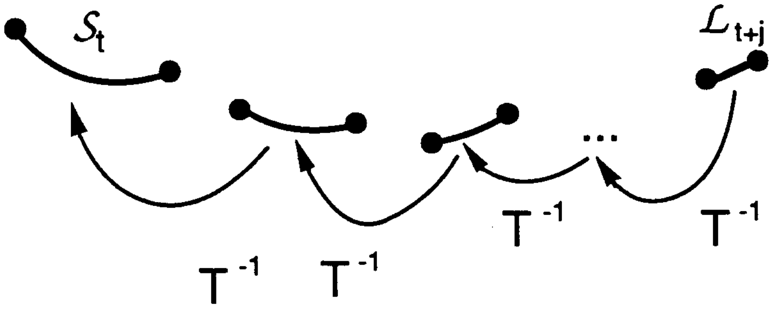
$$\lim_{t\to -\infty} |X_t - Z_t| = 0$$

Points which are both future and past asymptotic are homoclinic (to each other). If Y_0 is past asymptotic to X_0 and future asymptotic to Z_0 then it is heteroclinic from X_0 to Z_0.

If an orbit X_t is hyperbolic, then the set of points which are future or past asymptotic to X_0 form two smooth curves without self-intersection, crossing transversely at X_0, called the stable and unstable manifolds of X_0 (recall §2.3). 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 above, partial barriers from cantori, and minimizing heteroclinic 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 X_0 and Z_0 are future asymptotic, let $\mathcal{L}_j$, j >0, be the directed straight line segment from X_j to Z_j. Then the stable segment joining X_0 to Z_0 is

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



XBL 896-2135

Figure 7.10.1

Similarly a pair of past asymptotic points gives an unstable segment

$$\mathcal{U}_0 = \lim_{j\to\infty} T^j(\mathcal{L}_{-j})$$

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 $\{X_t\}$ and $\{Z_t\}$ be a future asymptotic pair. Denote the angle coordinate of X_t by x_t and of Z_t by z_t and the action difference by

$$\Delta F_t \equiv F(x_t, x_{t+1}) - F(z_t, z_{t+1})$$

Parameterize the stable segment $\mathcal{S}_t$ of a this future asymptotic by λ, such that $\mathcal{S}_t(0) = X_t$, $\mathcal{S}_t(1) = Z_t$ and $\mathcal{S}_{t+1}(\lambda) = T\,\mathcal{S}_t(\lambda)$. The area under $\mathcal{S}_t$, denoted A^s_t, is obtained by iterating the fundamental formula:

$$\begin{aligned} A^s_t &= A^s_{t+1} - \Delta F_t \\ &= A^s_{t+k} - \sum_{j=0}^{k-1} \Delta F_{t+j} \\ &= -\sum_{j=0}^{\infty} \Delta F_{t+j} \equiv \Delta W^f_t \end{aligned}$$

The last line follows because the area under the stable segment goes to zero as k→∞. The convergence of the sum for ΔW^f is guaranteed if the union of the two orbits is monotone. This is always the case whenever X_t and Z_t are minimizing orbits with the same rotation number (m,n), or are minimizing homoclinic orbits $(m,n)^+$ or $(m,n)^-$, or are minimizing with irrational rotation number, or one is minimizing and the other is an associated minimax orbit.

If $\{X_t\}$ and $\{Z_t\}$ are past asymptotic, and A^u_t is the area under their unstable segment $\mathcal{U}_t$, then a similar calculation gives

$$A^u_t = \sum_{j=-\infty}^{-1} \Delta F_{t+j} \equiv \Delta W^p_t$$

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

7.11 Flux–Homoclinic Orbits

We can combine the future and past sums if $\{X_t\}$ and $\{Z_t\}$ are homoclinic. The signed area between the unstable and stable segments, which is positive where $\mathcal{U}_t$ is above $\mathcal{S}_t$, is given by

$$A_t^s - A_t^u = \Delta W \equiv \sum_{j=-\infty}^{\infty} \Delta F_j$$

The sum certainly exists in the hyperbolic case. This equation defines the flux through a homoclinic pair of orbits, and the result is just Mather's ΔW. Note that it is independent of t because T is an area-preserving map, and hence the region contained between the stable and unstable segments has the same area for all time. For example we can let $\{X_t\}$ be a minimizing orbit $\{M_t\}$ and $\{Z_t\}$ be the corresponding minimax orbit $\{S_t\}$, corresponding to a cantorus or a partial separatrix. In this case the upward flux flowing between $\{M_t\}$ and $\{S_t\}$ is the difference in actions of these orbits.

7.12 Areas–Cantori and Separatrices

Once the area under a single segment is known, we obtain the total area under all the segments S_t after time t by simply summing:

$$\begin{aligned} A_t^f &= \sum_{k=1}^{\infty} A_{t+k}^s = -\sum_{k=1}^{\infty} \Delta W_{t+k}^f \\ &= -\sum_{k=1}^{\infty} k\Delta F_{t+k} \end{aligned}$$

This converges in the hyperbolic case, since ΔF_t approaches zero exponentially. Similarly the total area under all the unstable segments at time t and before is

$$\begin{aligned} A_t^p &= \sum_{k=-\infty}^{0} A_{t+k}^u = -\sum_{k=-\infty}^{0} \Delta W_{t+k}^p \\ &= -\sum_{k=-\infty}^{0} k\Delta F_{t+k} \end{aligned}$$

Consider the case of a homoclinic pair of orbits. Adding the above two results gives the area under the total partial barrier; however, in general the sum $A_t = A^p_t + A^f_t$ does not give a useful quantity because it depends on the choice of t. Indeed if we shift the index t by one in this formula we obtain $A_{t+1} = A_t + \Delta W$ where ΔW is the action difference between the homoclinic pair. Thus it is only when ΔW is zero for the homoclinic pair that a t-independent result is obtained. There are two important cases: cantori and separatrices.

To find the area under the partial barrier formed from a cantorus, suppose the cantorus has a single family of gaps, and let $\{X_t\} = \{X^l{}_t\}$ be the orbit of the left endpoints of a gap in the cantorus and $\{Z_t\} = \{X^r{}_t\}$ be the orbit of the right endpoints. These orbits are absolute minima of the action, and the action difference ΔW vanishes as proved by Mather. Backward iterates of the unstable segment of a gap and forward iterates of the stable segment form the cantorus partial barrier (§7.6). Recall that when the cantorus is hyperbolic, it has zero length (indeed zero dimension, see §6.4). This implies that the area under cantorus itself is zero. Thus total area under the cantorus partial barrier is given by the sum of the future and past areas under the stable and unstable segments alone:

$$A_\omega = -\sum_{t=-\infty}^{\infty} t\left[F(x^l_t,\, x^l_{t+1}) - F(x^r_t,\, x^r_{t+1})\right]$$

Note that the area under a partial barrier is independent of the construction of the partial barrier itself, depending only on the orbit of the gap endpoints.

Now we obtain the area under an upper partial separatrix for the simplest case of a (0,1) resonance around an orbit of period 1, or fixed point $x = x_F$. For this simple case we choose X_t to be the point $M^+{}_t$ of an upper minimizing homoclinic orbit and $Z_t = M^+{}_{t+1}$ to be the next point on the same orbit. The orbits $\{M^+{}_t\}$ and $\{M^+{}_{t+1}\}$ are homoclinic, since they are both homoclinic to the fixed point. Furthermore since they are on the same orbit, the action difference vanishes. The area under the complete upper partial separatrix is therefore

$$A^+_{(0,1)} = -\sum_{t=-\infty}^{\infty} t\left[F(m^+_t, m^+_{t+1}) - F(m^+_{t-1}, m^+_t)\right]$$
$$= \sum_{t=-\infty}^{\infty} \left[F(m^+_t, m^+_{t+1}) - F(x_F, x_F)\right]$$

The second formula for A^+ is be obtained from by subtracting the action of the fixed point, $F(x_F,x_F)$, from the first term in the brackets in while simultaneously adding it to the second. This is necessary to maintain convergence. Shifting t by one gives the second line.

For the lower partial separatrix the points of the homoclinic orbit move to the left, so that increasing t means decreasing x. Defining $\{X_t\}$ and $\{Z_t\}$ as successive points on the lower minimizing orbit we 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. We choose $X_t = M^+{}_t$ and $Z_t = M^+{}_{t+n}$. Again, since ΔW for these orbits is zero, the area under the complete upper partial separatrix is

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

The expression for A^- is identical except for signs.

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

$$A_{(m,n)} = A^+_{(m,n)} - A^-_{(m,n)}$$

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 cancelling, 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.

7.13 Chen's Area Formulae

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

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

Aubry 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 (Chen, 1987) has shown that these derivatives give the area functions:

$$A^{\pm}_{(m,n)} = \left.\frac{d^{\pm}L}{d\omega}\right|_{\omega=m/n}$$

$$A_{\omega} = \left.\frac{dL}{d\omega}\right|_{\omega \text{ irrational}}$$

Here the ± in the derivative indicates that the derivative is taken from the right or left, respectively. These formulae are obtained by constructing the derivatives as limits of the difference L(ω′) - L(ω) as ω′ approaches ω on minimizing periodic orbits, and showing that the result is given by our previous area formulae. For irrational frequency this formula give 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.

7.14 Area Devil's Staircase

The area function A_{ω} gives a devil's staircase: a function which is monotonically increasing, and has jumps at every rational value. When there are no invariant circles, and every cantorus is hyperbolic, Aubry has shown that this devil's staircase is complete, which means that the entire change in the function occurs across the rational jumps. Since these changes each represent the area of a resonance, this implies that the resonances fill phase space: all the area is taken up by resonances (MacKay, Meiss and Percival, 1987). This has profound implications for transport, as we will see in §8 (yet to be written!).

Appendix A: Differential Forms (Arnol'd, 1978)

A differential n-form is a functional which operates on n vectors to give a real number. A one form can be thought of as a covariant vector; it acts on contravariant vectors by the dot product. For example the form df is a covariant vector associated with the gradient of a function f; operating on a vector v with df gives the derivative of f in the direction of v:

$$df(v) = v^i \frac{\partial f}{\partial x^i}$$

A two form is a really an antisymmetric matrix. The form we will be most concerned with is the symplectic form $\omega = \Sigma dp^i \wedge dq^i$. Operating on a pair of vectors δz and $\delta z'$ with ω gives

$$\omega(\delta z, \delta z') = \delta z^i \omega_{ij} \delta z'^j$$

where ω is the symplectic matrix ($\omega = J^{-1}$). In the coordinates (q_i, p_i), $\omega = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}$. The value of $\omega(\delta z, \delta z')$ is the sum of the areas of the parallelapipeds formed by the vectors $(\delta q_i, \delta p_i)$ and $(\delta q'_i, \delta p'_i)$. These parallelapipeds are formed from the projections of the vectors δz and $\delta z'$ onto the N canonical planes; each area is signed.

Differential n-forms can be integrated over n dimensional surfaces. For example, choose an arbitrary two dimensional surface $\mathcal{C}$ embedded in the 2N dimensional manifold $\mathcal{S}$. Associate an orientation to $\mathcal{C}$ by choosing a direction to traverse the boundary of $\mathcal{C}$. Preservation of ω means that the integral

$$A = \sum_{i=1}^{N} \int_{\mathcal{C}} dp_i \wedge dq_i$$

is conserved along the flow. This integral is a sum over the projected areas of the surface $\mathcal{C}$, the wedge product means that the areas are signed according to whether the boundary of the projection is traversed in a counterclockwise or a clockwise direction.

Appendix B: Circle Maps

Here we review a few basic facts about homeomorphisms of the circle (See pp. 73-95 of (Cornfeld, Fomin and Sinai, 1982)). Let $\alpha(x)$ be a continuous, monotonic, increasing function of x satisfying $\alpha(x+1) = \alpha(x)+1$.

B1 Lemma: There exists an ω such that for all $x \in R$ and (m,n)

$$\begin{aligned} n\omega > m \quad &\Rightarrow \quad \alpha^n(x) - m > x \\ n\omega < m \quad &\Rightarrow \quad \alpha^n(x) - m < x\,. \end{aligned}$$

The ω which satisfies this lemma is the rotation number. 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

$$|\alpha^n(x) - x - n\omega| \leq 1$$

This result implies

B2 Theorem : The limit

$$\omega = \lim_{n\to\infty} \frac{\alpha^n(x)}{n}$$

exists and does not depend on the choice of $x \in \mathcal{R}$. 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 P be the set of limit points of the orbit: $x \in P$ if there is a sequence $x_j = \alpha^j(x_0)$ such that $x_j \to x$ as $j \to \infty$ .By definition P is closed. Then

B3 Theorem (Denjoy, 1932; Poincaré, 1885): When ω is irrational, then

a) P is independent of the choice of x_0;

b) P is invariant;

c) either P is the entire circle, or P is a Cantor set.

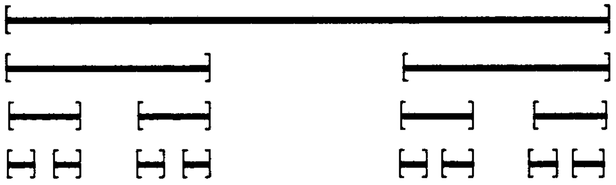
A Cantor set C is a non-empty, perfect, totally disconnected, compact set:

perfect $\Leftrightarrow$ Every point in the set is a limit point of other points in the set: $\forall\, x \in C, \exists$ a sequence $x^{(n)} \in C$ such that $x^{(n)} \neq x$ and $x^{(n)} \to x$ as $n \to \infty$

totally disconnected $\Leftrightarrow \forall\, x,y \in C\ \exists$ A and B, $x \in A$ and $y \in B$, such that $C = A \cup B$ and $A \cap B = \varnothing$.

compact $\Leftrightarrow$ for every open cover of C there is a finite sub cover.

The standard example of a Cantor set is a subset of the interval [0,1]. Remove the open interval (1/3,2/3), leaving two closed intervals. Remove the middle third from each of these. Continue this procedure *ad infinitum.*



XBL 896-2158

Figure B.1

Note however, that the definition of the Cantor set is a purely topological one, it does not require the set to be embedded in any other space.

A Cantor set contains an uncountable number of points, in particular there are points which are not at the end-point of any gap. To see this consider a coding of the points in the set in a base three representation: points in the set have no 1's in their base three "decimal"expansion. On the other hand, points which lie on the end points of a gap have base three expansions which truncate. There are an uncountable number of non-truncating sequences of 0's and 2's.

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 typically have zero Hausdorff dimension.

The complement of a Cantor set is a countable set of gaps.

References

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

Aubry, S. (1983), "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.

Aubry, S. (1978), On the Dynamics of Structural Phase Transitions, Lattice Locking, and Ergodic Theory, Solitons and Condensed Matter, (Springer-Verlag, New York).

Aubry, S. (1983), "Exact Models with a Complete Devil's Staircase," J. Phys. C, **16**: 2497-2508.

Berry, M. V. (1981), "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.

Cary, J. R. and R. G. Littlejohn (1983), "Noncanonical Hamiltonian Mechanics and Its Application to Magnetic Field Line Flow," Ann. Phys., **151**: 1-34.

Chen, Q. (1987), "Area as a Devil's Staircase in Twist Maps," Phys. Lett., **123A**: 444.

Cornfeld, I. P., S. V. Fomin and Y. G. Sinai (1982), Ergodic Theory, Grundlehren der mathematishcen Wissenschaften, (Springer-Verlag, New York).

Denjoy, A. (1932), "Sur les Courbes Défines par les Équations Différentielles a la Surface du Tore," J. Math. Pures et Appl., **11**: 333-375.

Gelfand, I. M. and S. V. Fomin (1963), Calculus of Variations, (Prentice Hall, Englewood Cliffs, NJ).

Goroff, D. L. (1985), "Hyperbolic Sets for Twist Maps," Erg. Th. Dyn. Sys., **5**: 337-339.

Greene, J. M. (1968), "Two-Dimensional Area Preserving Mappings," J. Math. Phys., **9**: 760-768.

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 Transistion of an Incommensurate Structure," Phys. Rev. A, **36**: 5858-5861.

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.

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. (1985), "Sur les Courbes Invariantes par les Diffeomorphismes de L'anneau. Vol 2," Asterisque, :

Herman, M. R. (1988),"Existence et Non-existence de Tores Invariants par des Diffeomorphismes Symplectiques," Ecole Polytechnique .

Katok, A. (1982), "Some Remarks on the Birkhoff and Mather Twist Map Theorems," Ergod. Th. & Dynam. Sys., **2**: 185-194.

Ketoja, J. A. and R. S. MacKay (1988),"Fractal Boundary for the Existence of Invariant Circles for Area-Preserving Maps: Observations and Renormalization Explanation," Univ. of Warwick .

Khinchin, A. Y. (1964), Continued Fractions, (University of Chicago Press, Chicago).

Kook, H. T. and J. D. Meiss (1988), "Periodic Orbits for Reversible, Symplectic Mappings," Physica D, **in press**:

Lanford, O. E. (1973), Introduction to the mathematical theory of dynamical systems, Chaotic Behavior of deterministic systems, (North Holland, Amsterdam).

Lichtenberg, A. J. and M. A. Lieberman (1982), Regular and Stochastic Motion, (Springer-Verlag, New York).

MacKay, R. S. (1988), "Hyperbolic Cantori have Dimension Zero," J. Phys. A., **to appear**:

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., J. D. Meiss and I. I. C. Perciva (1984), "Stochasticity and Transport in Hamiltonian Systems," Phys. Rev. Lett., **52**: 697-700.

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 J. Stark (1989),"Converse KAM Theory for Symplectic Twist Maps," University of Warwick .

MacKay, R. S. and I. C. Percival (1985), "Converse KAM: Theory and Practice," Comm. 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 .

MacKay, R. S., J. D. Meiss and I. C. Percival (1987), "Resonances in Area Preserving Maps," Physica D, **27**: 1-20.

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," Erg. Theory and Dyn. Sys., **2**: 301-309.

Mather, J. N. (1986), "A Criterion for Non-existence of Invariant Circles," Publ. Math. I.H.E.S., **63**: 153-204.

Meiss, J. D. (1986), "Class Renormalization: Islands around Islands," Phys. Rev. A, **34**: 2375-2383.

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 Univ. Press, Princeton).

Percival, I. C. (1979), Variational Principles for Invariant Tori and Cantori, Nonlinear Dynamics and the Beam-Bean Interaction, (Amer. Inst. Phys., New York).

Poincaré, H. (1885), "Mémoire sur les Courbes Définies par une Équation Différentielle, III," J. Math. Pures et Appl., **1**(4): 167-244.

Stark, J. (1986), On Invariant Circles for Area-Preserving Maps, (Ph.D. Thesis, University of Warwick).

Umberger, D. K. and J. D. Farmer (1985), "Fat Fractals on the Energy Surface," Phys. Rev. Lett., **55**: 661-664.

Wojtkowski, M. (1981), "A Model Problem with the Coexistence of Stochastic and Integrable Behaviour," Commun. Math. Phys., **80**: 453-464.

LAWRENCE BERKELEY LABORATORY
TECHNICAL INFORMATION DEPARTMENT
1 CYCLOTRON ROAD
BERKELEY, CALIFORNIA 94720