TRANSPORT IN HAMILTONIAN SYSTEMS

R.S.MacKay[1], J.D.Meiss[2] and I.C.Percival,

Department of Applied Mathematics,

Queen Mary College, London E1 4NS

Revised December 1983, to be published in Physica D

ABSTRACT

We develop a theory of transport in Hamiltonian systems in the context of iteration of area preserving maps. Invariant closed curves present complete barriers to transport, but in regions without such curves there are still invariant Cantor sets named cantori, which appear to form partial barriers. The flux through the gaps of the cantori is given by Mather's differences in action. This gives useful bounds on transport between regions, and for one parameter families of maps it provides a universal scaling law when a curve has just broken. The bounds and scaling law both agree well with numerical experiments of Chirikov and help to explain an apparent disagreement with results of Greene. By dividing the largest irregular component of phase space into regions separated by the strongest partial barriers, and assuming that the motion is mixing within these regions, we present a global picture of transport, and indicate how it can be used, for example, to predict confinement times and to explain longtime tails in the decay of correlations.

[1] Address from 1 April 1984: Mathematics Institute, University of Warwick, Coventry CV4 7AL, England.

[2] Permanent address: Institute for Fusion Studies, University of Texas, Austin TX 78712 USA.

CONTENTS

1. INTRODUCTION

The motion of Hamiltonian systems is typically neither entirely regular, nor entirely irregular, but the phase space consists of a complicated mixture of regular and irregular components. In the regular components the motion is quasiperiodic and orbits lie on tori or 'KAM surfaces', while in the irregular components the motion appears stochastic or chaotic and seems to fill regions of higher dimension. In this paper we study transport in the irregular components and initiate a theory of the organisation inherent in the apparent chaos. This is achieved by recognising a natural division of the irregular components into regions separated by partial barriers formed by joining the gaps in invariant Cantor sets. We discuss the mechanism of transport between these regions. The theory is not yet complete, but we believe that the ideas introduced here are of central importance for a broad range of systems that are neither completely ordered nor completely chaotic.

The theory of transport in Hamiltonian systems has many applications, including calculations of particle loss from plasmas and accelerators, rates of chemical reactions, wave heating rates in plasmas, and similar quantities in other fields.

The simplest Hamiltonian systems with an irregular component have two degrees of freedom and by using surfaces of section they can be reduced to the iteration of area preserving maps. The theory in this paper is formulated in terms of maps, but all the concepts and methods apply to the corresponding Hamiltonian systems, without the need to carry out the reduction explicitly.

Figure 1.1 shows an example of an irregular component,

produced by following one orbit for a long time. The motion in an irregular component, however, need not be completely chaotic. Orbits can take a long time to explore the region accessible to them. In figure 1.2, orbits were started in the upper rectangle and removed if they entered the lower square. From the sudden changes in density, there appear to be partial barriers to the motion of the phase points.

What is limiting the transport in irregular components?

We propose that the main obstructions are invariant Cantor sets, known as cantori. They are invariant sets in which the motion has irrational frequencies, and though they resemble invariant circles, they have an infinite number of gaps in them, as shown in figure 1.3. Two are shown in figure 1.2 (or strictly, orbits homoclinic to cantori), and are clearly presenting barriers to the irregular orbit. Their existence was proposed by Percival [1] and Aubry [2]. They gave an explicit example, and a general proof of their existence has been given by Aubry et al [3], Mather [4] and Katok [5].

Most of the gaps in a cantorus are very small, so that their total length is finite. Even when there are large gaps, orbits can take a long time to get through. For one parameter families of maps and for the near critical parameter values where an invariant circle has just broken up, the gaps become very small, so the cantori certainly present the strongest partial barriers.

We want some way of determining the flux through the gaps. To define the flux we close the gaps to form a closed curve. We are thus led to the problem of minimising the flux across closed curves. This idea goes back at least to Wigner [6], and similar ideas are being developed by Bensimon and Kadanoff [7].

We make a sharp distinction between flux, which represents the area crossing a closed curve in one iteration of the map, and transport, which refers to the combined long term effects of the flux across many barriers. So we divide the body of the paper into two parts. The first part, consisting of sections 2-6 is a local theory of flux across a closed curve. The second part, consisting of sections 7-11 initiates a global theory of transport.

In section 2 we define the flux, and introduce partial barriers. The flux across partial barriers passes through 'turnstiles', which play a central role in the theory. The construction from cantori of partial barriers with turnstiles is described and illustrated. In section 3 the theory is put on a firmer foundation with the introduction of stationary action principles, and in section 4 it is shown that the flux across partial barriers is given by differences in action ΔW, that Mather introduced for cantori [8]. The properties of cantori and partial barriers are related to the properties of periodic orbits. Limiting properties of the ΔW for the periodic orbits provide strong evidence that the cantori with noble frequency are the most important partial barriers. The relation between periodic orbits and cantori also provides a practical method of finding cantori, constructing approximate partial barriers, and evaluating the flux across them. In section 6, this is illustrated for the example of the standard map, introduced in section 5.

In section 7 we introduce a global theory of transport that is radically different from previous theories. The properties of long term transport are derived from the properties of the flux on the basis of a Markov model and a related circuit model. The phase space is divided into separate irregular components by

invariant circles forming complete barriers. Each irregular component is itself subdivided into regions by partial barriers with turnstiles formed from the more important cantori. In the Markov model transition probabilities between these regions are determined by ratios of areas of turnstiles to areas of regions. In the circuit model the successive transitions are replaced by a continuous flow of charge between capacitances to ground, representing the regions, through resistors representing the turnstiles.

Explanations are given of multiple time constants and power laws for decay in section 8, and formulae for the mean transit time from one region of an irregular component to another are presented and compared with numerical experiment in section 9. In section 10 we discuss shortcomings of the model, and the sort of modifications that will have to be made. In section 11 we obtain a scaling law for near critical parameter values of one parameter families of maps. This resolves an apparent discrepancy between results of Chirikov [9] and Greene [10].

Our studies show that there can be considerable order in irregular motion. This is particularly true near cantori with small gaps, like those near the invariant circles that form the boundaries of irregular regions. The partial barriers define an approximate or adiabatic invariant. Where the turnstiles are aligned it is possible to consider them together as a chimney. Motion across the associated partial barriers is then restricted to this chimney, where there is transport in both directions. Elsewhere the motion resembles that on invariant circles, being interrupted only when the phase point arrives in a turnstile. It is possible, however, that a better adiabatic invariant may be obtained by a choice of partial barriers which shares out the flux

between all the gaps.

2. FLUX, BARRIERS AND TURNSTILES

(a) Definition of flux

An iterated area preserving map in the plane

$$(x',p') = T(x,p) \tag{2.1}$$

is the discrete time equivalent of an area preserving flow defined by Hamilton's equations of motion for the phase point of a system of one degree of freedom. A consistent definition of flux across a closed curve C in the plane is given by the area occupied by all the phase points mapped from the interior R to the exterior by one iteration of the map. Because the area is preserved this is equal to the area mapped from the exterior into R.

Let

$$C' = T(C), \quad R' = T(R) \tag{2.2}$$

be the images of C and R, so that R' is the interior of C'. Then

$$\text{flux across } C = \text{Ar}(R'-R) = \text{Ar}(R-R') \tag{2.3}$$

where $\text{Ar}(\ )$ represents an area and $R'-R$, $R-R'$ are the shaded regions of figure 2.1.

A closed curve is topologically a circle, and we use that name on the understanding that it is only exceptionally a geometrical circle. If C is an invariant circle then $C = C'$,

R = R' and the flux is zero: it is a complete barrier to the motion of the phase points induced by the map T.

The same definition and results apply if C is a chain of circles, as illustrated in figure 2.2. In this case the image of each curve may be shifted as shown in the figure. One can choose C and C' to be the same, however, except for one link in the chain. If C is an invariant chain of circles around the phase points of a stable periodic orbit, then the flux is zero.

(b) Cylindrical phase space

When the variable $2\pi x$ represents a physical angle, then x and x+1 represent the same configuration. Then the map is periodic in the plane:

$$(x',p') = T(x,p) \quad ===> \quad (x'+1,p') = T(x+1,p) \tag{2.4}$$

and can be represented without ambiguity on a cylinder, with $2\pi x$ as its angular configuration coordinate.

A phase point that passes once around the circle p=0 corresponds to a physical rotation, and so do any circles obtainable from it by continuous transformation on the cylinder. Each such 'rotational' circle C divides the phase space into two infinite parts. Let R be the region below the circle C, and R' the region below C'=T(C). If the regions are infinite, then in general

$$Ar(R'-R) - Ar(R-R') \neq 0 \tag{2.5}$$

so the nett flux is not zero and the upward and downward fluxes are not the same. By considering the area of the region between any two rotational circles, it is easy to see that the nett flux is independent of C; this flux is sometimes known as the Calabi invariant.

But if the area on one side is finite, as when p is a radial coordinate from a fixed point, then the fluxes must be equal, so the nett flux is zero.

(c) Partial barriers and turnstiles

Invariant circles present complete barriers to the motion. But there may also be partial barriers; sometimes many iterations are required to cross them. In later sections we will show that a particularly important type of partial barrier has the form illustrated in figure 2.3. C and C' are rotational circles and

$$C'=T(C) \qquad (2.6)$$

where T is a twist map with zero nett flux. Although neither C nor C' is invariant, they coincide over much of their length, between A and B, and between D and A. The curve DAB is a partial barrier. In the gap between B and D, the circles C and C' must cross at least once, otherwise the upward or the downward flux would be zero. Many crossings are possible, but make no essential difference, so we suppose that they cross only once, at the point E.

As before the upward and downward flux are the same. All phase points that pass upwards through the partial barrier must go through R'-R and all those that pass downwards must go through

R-R', so the whole structure in the gap between B and D acts like a revolving door or turnstile.

Partial barriers with turnstiles are our principal concern.

(d) Minimum flux and cantori

In 1937 Wigner [6] formulated a variational principle for transport properies of systems satisfying Hamilton's equations, in which the more important surfaces were determined by minimizing the flux across them. Keck and his collaborators have developed and used this principle to estimate rates of reactions [11].

Similar ideas are being developed by Bensimon and Kadanoff [7]. The idea of 'almost invariant curves' also occurs in theories of transport across broken separatrices, e.g. [15-17].

We are not able to prove that partial barriers with turnstiles formed from cantori with noble frequency minimize the flux, but the numerical evidence is very strong. The barriers with turnstiles are constructed from the stable and unstable manifolds of the points at the edge of one of the gaps, usually chosen to be the largest. Ignoring some complications these manifolds are obtained as follows.

A 'forward orbit' from the point (x,p) is the ordered sequence of phase points

$$T_t(x,p) \qquad (0 \leqslant t \leqslant \infty) \qquad (2.7)$$

where t is an integer time variable that labels the iterations of the map. Similarly for a backward orbit, but with $-\infty \leqslant t \leqslant 0$.

Take any point x_0 in the gap. For small enough values

of p the forward orbit from (x_0,p) will tend to be below the cantorus, and for large enough values it will tend to be above, so it is reasonable to assume that for some value p_0 of p the forward orbit from (x_0,p) converges to the cantorus. The curve formed from all the points (x_0,p_0) is continuous and closes the chosen gap. The forward orbits from these points close all the forward image gaps $T^t(\text{gap})$, with positive t. A similar curve gives a set of backward orbits closing the backward image gaps $T^t(\text{gap})$ with negative t. All the gaps are then closed by a single curve joining their end points, except the chosen gap, which contains a turnstile. An example will be illustrated in figure 6.2.

The curves forming the turnstile need not be given by single-valued functions $p=p_\pm(x)$. In fact it is easier to obtain them by the more stable procedure of joining the ends of a forward image gap with large t by a straight line in the phase plane, and then iterating backwards to the chosen gap. Similarly for some backward image gap. More details of the methods used to obtain the figure are given in section 6.

There are two types of complication. One is that the turnstile may have more than two lobes. The other is that the process described may not exhaust all the gaps. If it does not, it is necessary to have more than one chosen gap in order to generate the others by forward and backward iteration.

To obtain the flux one needs an easy way of calculating the area of a turnstile without finding the stable and unstable manifolds. This is provided by the principle of stationary action.

3. ACTION PRINCIPLES

It is well known that the dynamics of Hamiltonian systems can be formulated variationally, using the principle of stationary action. The aim of the next two sections is to show that this is intimately related to questions of flux. We review the action principle for area preserving maps, and sketch how it leads to existence of various types of orbit and invariant set, including cantori. We show that the area of a turnstile in a cantorus is precisely the difference in action between the orbits on the cantorus and an orbit homoclinic to it, passing through the pivot of the turnstile. Evaluation of limits of differences of action between periodic orbits gives a practical method of finding the area of turnstiles for cantori, justified by continuity results of Mather. Numerical evidence suggests that the local minima of ΔW are at noble frequencies, indicating that noble cantori are the strongest barriers.

An area preserving map $(x',p')=T(x,p)$ is a special case of a canonical or symplectic transformation, so if it satisfies the twist condition

$$\partial x'/\partial p \neq 0 \qquad \text{(all } x,p) \qquad (3.1)$$

then there exists a generating function $F(x,x')$ such that

$$p = -\partial F(x,x')/\partial x, \qquad p'=\partial F(x,x')/\partial x' \qquad (3.2)$$

It has a geometrical interpretation as the area under the image of the vertical line through x, between an arbitrary reference line and the vertical through x' (see figure 3.1).

Note from (3.1) and (3.2) that:

$$\frac{\partial^2 F}{\partial x \partial x'} \neq 0 \qquad (3.3)$$

Conversely, if (3.3) is satisfied then the relations (3.2) can be inverted to define an area preserving map that satisfies the twist condition. Generating functions are unique up to addition of a constant.

For our purposes a 'twist map' is an area preserving map that satisfies the twist condion and, if defined on a cylinder, has zero nett flux. The value of $\partial x'/\partial p$ always has the same sign. We adopt the convention that it is positive, so the second derivative (3.3) is always negative.

For example, the motion of an otherwise free particle of unit mass subject to position dependent impulses $P(x)$ at unit intervals of time can be represented by a twist map

$$p' = p + P(x), \qquad x' = x + p' \qquad (3.4)$$

with generating function

$$F(x,x') = 1/2\ (x-x')^2 - V(x) \qquad (3.5)$$

where $V(x)$ is the potential for the function $P(x)$:

$$P(x) = -dV(x)/dx \qquad (3.6)$$

The generating function (3.5) is the discrete time equivalent of the Lagrangian for a free particle with potential $V(x)$.

The Lagrangian for a continuous time system of one freedom

can be used in a stationary action principle for the motion. For discrete time there are two distinct principles, one for sums over successive points of an orbit, the other for an integral over an invariant set of the motion.

Consider the first. It follows from (3.2) that if x,x',x'' are three successive values of x on an orbit, then

$$(\partial/\partial x')[F(x,x')+F(x',x'')] = 0 \qquad (3.7)$$

and conversely.

For integers r and s with $r+1<s$ let

$$x_t \qquad (r\leqslant t\leqslant s) \qquad (3.8)$$

be an arbitrary sequence of real values of x subject to fixed initial x_r and final x_s. The variable t is an integer time. From (3.7) this sequence defines a segment of a T-orbit if and only if the action sum

$$W_{rs} = \sum_{t=r}^{s-1} F(x_t,x_{t+1}) \qquad (3.9)$$

is stationary with respect to arbitrary variations of the intermediate x_t.

An infinite sequence defines an orbit if and only if every finite segment has stationary action. This applies to forward orbits, backward orbits and complete orbits (section 2).

We are particularly interested in maps where $2\pi x$ is an angle. So for the rest of this section and the next one, consider

area preserving twist maps on the cylinder, and the corresponding periodic maps on the plane. If $F(x,x')$ generates a periodic area preserving twist map, then clearly $F(x+1,x'+1)$ generates the same map. Thus they can differ only by a constant. In fact that constant is precisely the nett flux. To see this take any curve C joining (x,x') to $(x+1,x'+1)$. It defines a circle $p(x)$ and its image $p'(x')$ on the cylinder, by (3.2). The difference in areas under them is:

$$\int p'(x')dx' - \int p(x)dx = \int \frac{\partial F}{\partial x'}(x,x')dx' + \frac{\partial F}{\partial x}(x,x')dx$$
$$= \Big[F(x,x')\Big]_C \qquad (3.10)$$

The assumption that the nett flux is zero is essential for the existence theorems that follow, though not for the interpretation of differences in actions.

With our sign convention, an essential role in periodic area preserving twist maps with zero nett flux is played by orbits which minimise the action of each finite segment. Aubry and le Daeron [3] showed that minimising orbits satisfy an order preserving property, viz.

$$x_t + j < x_{t'} \quad \text{iff} \quad x_{t+1} + j < x_{t'+1} \qquad (3.11)$$

for integers j,t,t'. It follows that they have a rotation number, $\lim x_t/t$. This may be rational or irrational. We sketch the derivation of three types of minimising orbit.

(i) Minimising periodic orbits

From the action principle, a sequence (x_t) with

$$x_{t+n} = x_t + m \qquad (3.12)$$

gives a periodic orbit (of rotation number m/n) iff its action:

$$W = \sum_{t=0}^{n-1} F(x_t, x_{t+1}) \qquad (3.13)$$

is stationary with respect to variations keeping $x_n = x_0 + m$. W is bounded below, and hence has a minimum. Aubry [3] and Mather [8] proved that any minimum gives a periodic orbit.

The linear stability of a periodic orbit is given by the eigenvalues of the tangent map round the orbit. For an area preserving map, their product is 1, so it is sufficient to specify the residue [10]:

$$R = (2-\lambda-1/\lambda)/4 \qquad (3.14)$$

where λ is either eigenvalue. In [12] it is shown that minimising orbits have non-positive residue, i.e. real and positive eigenvalues, so they are unstable orbits.

(ii) Minimising quasiperiodic orbits

The variational principle (3.13) for periodic orbits was generalised by Percival [13] to quasiperiodic orbits, of rotation number ν. Instead of a sum one makes an integral stationary, viz.

$$W = \int_0^1 d\theta\ F(x(\theta), x(\theta+\nu)) \qquad (3.15)$$

over increasing functions $x(\theta)$, satisfying $x(\theta+1) = x(\theta) + 1$. Mather [4] proved existence of a minimising $x(\theta)$. It gives rise to an invariant set parametrised by θ, and $p(\theta)$ is defined by:

$$p(\theta) = -F_1(x(\theta), x(\theta+\nu)) \qquad (3.16)$$

where subscript 1 refers to the derivative with respect to the first argument. When projected back onto the cylinder, it may be a circle, or it may have a dense set of gaps, leaving a Cantor set. The latter case is what we call a cantorus. In either case, it consists of minimising orbits of rotation number ω, and the induced map on θ is simply a uniform rotation:

$$\theta' = \theta + \nu \qquad (3.17)$$

This invariant set can also be derived as the limit as $m/n \to \nu$ of minimising periodic orbits of rotation number m/n [3,5].

In the case that the set is a circle, the value of W is proportional to the area under the circle. Probably there is a similar interpretation in the case of a cantorus, but we have not worked it out.

(iii) Minimising heteroclinic orbits

The limit of minimising orbits of rotation number ν, as $\nu \to m/n$ from above, gives a minimising orbit of rotation number m/n. It is not periodic, however, but converges to one minimising periodic orbit of rotation number m/n in forward time and another, or the same one with its points relabelled, in backward time, sliding to the right. Similarly the limit as $\nu \to m/n$ from below gives a minimising heteroclinic orbit sliding to the left.

4. FLUX AND DIFFERENCES IN ACTION

For each minimising orbit, except when the minimising set is a circle, one can prove existence of a companion "minimax" orbit, corresponding to a saddle in the action [14,8]. This gives three types of minimax orbit. In each case, the difference between the minimax and the minimum action can be interpreted as a flux.

(i) Minimax periodic orbits

Given a minimising periodic orbit l_t ("l" for left), define r_t ("r" for right) to be the same orbit but with the points relabelled so that r_t is the closest point to the right of l_t. Then r_t has the same action (3.13) (provided the nett flux is zero), so there must be a saddle in between. This gives another periodic orbit c_t ("c" for centre) of the same rotation number [8]. In [12] it is shown that it has non-negative residue, i.e. the eigenvalues are a complex conjugate pair on the unit circle or a reciprocal pair of negative reals.

Taking the lowest saddle, Mather defines $\Delta W_{m/n}$ to be the difference in actions. It can be interpreted as the area that is transported between the minimising and minimax periodic orbits per iteration. To see this, join up the gap between two neighbouring periodic points of a minimising periodic orbit, by any curve C $p_0(x_0)$ passing through a point of the minimax orbit (figure 4.1). Its n^{th} iterate also forms such a curve. Except in the chosen gap, the images $p_t(x_t)$ of C, with t=1 to q, form an invariant curve, a barrier to transport. But at each iteration the area marked 1 is mapped somewhere above the barrier, and the area marked 2 is mapped below. So the figure of 8 acts like a revolving door or turnstile.

Now the area of each side of the turnstile is precisely $\Delta W_{m/n}$ (algebraically), because the area of 1, for example, is:

$$\int_{l_n}^{c_n} p_n dx_n - \int_{l_0}^{c_0} p_0 dx_0 = \int \frac{\partial F}{\partial x_n}(x_{n-1}, x_n) dx_n + \int \frac{\partial F}{\partial x_0}(x_0, x_1) dx_0$$

Add in $\sum_{t=1}^{n-1} \int dx_t \ \frac{\partial F}{\partial x_t}(x_{t-1}, x_t) + \frac{\partial F}{\partial x_t}(x_t, x_{t+1})$

which is zero by (3.7). Hence the area is

$$\Delta W = \left[\sum_{t=0}^{n-1} F(x_t, x_{t+1}) \right]_{(l_t)}^{(c_t)} \qquad (4.1)$$

which is the difference in actions between (c_t) and (l_t). Similarly, the area of 2 is the difference between the actions of (r_t) and (c_t), which is equal and opposite to ΔW.

One caution is that one really wants the geometric area transported, not the algebraic area. $T^n C$ could cross C more than once, giving more transport than ΔW (compare section 2). This would happen if there were more than one minimising and one minimax periodic orbit of given rotation number. One would have to add all the differences in actions between neighbours.

Figure 4.2 shows $\Delta W_{m/n}$ for a selection of rationals for a typical map (actually it is for a very special map, the universal map for the neighbourhood of a critical noble circle [30], but being universal it is also typical!). They are plotted against the position p of the minimax point on a certain symmetry line. Some of the points are labelled by their rotation number in continued fraction notation. The continued fraction expansion of an irrational number ν is the unique sequence $[a_0, a_1 \ldots]$ of integers, $a_n \geq 1$ for $n \geq 1$, such that

$$\nu = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ldots}} \qquad (4.2)$$

For rationals, the continued fraction expansion must terminate, but there are always two possible terminations, since

$$[a_0,\ldots a_k+1] = [a_0,\ldots a_k,1] \qquad (4.3)$$

This ambiguity is used to connect the points of figure 4.2 in a binary tree. We make the following two observations.

$\Delta W_{m/n}$ always appears to decrease as one follows the tree. This makes sense because after n iterations the region that passed up through the turnstile for rotation number m/n intersects the turnstile again. So some of it will go back down. Thus, for small n, $\Delta W_{m/n}$ is an overestimate of the long term flux.

$\Delta W_{m/n}$ appears to go to a limit as one follows each route down the tree. We could label the limits as ΔW_ν, $\nu=[a_0,\ldots]$ for branching sequences which take a_n steps between successive changes of direction. The case of termination with some $a_n=\infty$ is also allowed for sequences which eventually stick to the left or right hand direction. In fact, Mather [8] has proved that $\Delta W_{m/n}$ does go to limits as m/n goes to irrationals and monotonically to rationals. Furthermore the limits correspond to differences in action between aperiodic minimax orbits and the minimising ones.

(ii) Minimax quasiperiodic orbits

If the minimising set of rotation number ν is a circle, then Mather proved that $\Delta W_{m/n}\to 0$ as $m/n\to\nu$. There is no flux across it, and no minimax quasiperiodic orbit. Suppose instead that we have a cantorus. Its projection on the x-coordinate is the complement of a dense set of gaps. They fall into families under the relation of being images under T. Choose a family, and one particular gap (l_0,r_0) of the family. Let l_t and r_t be

the orbits of the endpoints of this gap. Then, since the total length of the gaps is finite, the sum:

$$\Delta W(x_t) = \sum_{t=-\infty}^{\infty} F(x_t, x_{t+1}) - F(l_t, l_{t+1}) \tag{4.4}$$

converges for sequences x_t lying in the orbit of the gap, i.e.

$$l_t \leqslant x_t \leqslant r_t \tag{4.5}$$

Furthermore, it is non-negative, and is zero for the sequences l_t and r_t. Thus there is a minimax orbit c_t in between. Since it lives in the orbit of the initial gap, and the gap widths go to zero, it is homoclinic to the Cantor set, i.e. converges to it in both directions of time.

Let ΔW_ν be the minimax value. It can be interpreted as the flux through that family of gaps in the Cantor set. As in the periodic case, we can blame all the transport on one gap. Define the stable set $C^+(x,p)$ to be the set of points (x',p') such that the distance between $T^t(x',p')$ and $T^t(x,p)$ goes to zero as $t \to +\infty$. Similarly define the unstable set $C^-(x,p)$ of (x,p). Choose a gap. Both endpoints have the same stable set, and the same unstable set, since they converge together in both directions of time. Also the minimax orbit belongs to both sets. Thus one expects a picture like figure 4.3. Let us describe the curves in the gap by functions $p^+(x_0)$ and $p^-(x_0)$, which may not be singlevalued, however. Then we can join up the forward images of the gap with the images of C^+ and the backward images with the images of C^-, forming a barrier with turnstile. Now the area of the turnstile is:

$$\int_{l_0}^{c_0} dx_0 \; p^-(x_0) - p^+(x_0)$$

$$= \int dx_0 \; \frac{\partial F}{\partial x_0}(x_{-1}, x_0) + \frac{\partial F}{\partial x_0}(x_0, x_1)$$

Now add in $\sum_{\substack{t=-\infty \\ t\neq 0}}^{\infty} \int dx_t \frac{\partial F}{\partial x_t}(x_{t-1},x_t) + \frac{\partial F}{\partial x_t}(x_t,x_{t+1})$

which is zero by stationarity, to get:

$$\text{Area} = \left[\sum_{t=-\infty}^{\infty} F(x_t,x_{t+1}) \right]_1^c = \Delta W \qquad (4.6)$$

If there is more than one family of gaps, then the total flux through the cantorus is given by the sum of the ΔWs for the families.

(iii) Minimax heteroclinic orbits

In the same way that there are minimax orbits between endpoint orbits of a minimising Cantor set, there are minimax orbits between the minimising heteroclinic orbits and their translates. The minimax value of ΔW can be interpreted as the area of a turnstile in the broken separatrix of an island chain as shown in Fig.4.4. In fact, this picture of flux across broken separatrices via a turnstile already appears in references such as [15-17], though not the interpretation of ΔW.

Which of these three types of barrier with turnstile provides the strongest barrier? Clearly, the answer is the ones with the smallest ΔW. Figure 4.2 suggests that the smallest ΔW in any interval of rotation number is attained by alternating direction at each step along the tree. This is the prescription for approaching a noble rotation number, i.e. one for which there exists an integer N such that $a_n=1$ for $n\geq N$. So it suggests that the noble cantori form the greatest barriers to transport, and we should regard them as the organising centres for the problem. In particular, they are more important than broken separatrices. All

the nearby ΔWs will be roughly the same, by continuity of ΔW at irrationals, but where the turnstiles overlap significantly, the transport is limited mainly by the smallest one. Not all nobles will be equally important, of course. From the figure, the more important nobles are those with the smaller values of the integer N defined above.

In conclusion, the action principle gives a practical way of evaluating the area of turnstiles in cantori. The easiest way is to find minimising and minimax periodic orbits of rotation number $m/n \to \nu$, and take lim $\Delta W_{m/n}$

5. THE STANDARD MAP

(a) General properties

Most one parameter families of maps have similar qualitative features. In this section barriers and turnstiles are constructed for a particular family which has all these features:

$$p' = p - (k/2\pi)\sin(2\pi x);$$

$$x' = x + p'. \qquad (5.1)$$

This is the Chirikov-Taylor or standard map. The variable $2\pi p$ may be regarded as the angular momentum of a free rotor subject to angle dependent impulses of strength k at unit intervals of time. The standard map is a twist map of the form (3.4) and has a variational principle (3.5) with

$$V(x) = -(k/4\pi^2)\cos(2\pi x) \qquad (5.2)$$

By using the invariance with respect to unit horizontal and vertical translations and reflection through the origin, we can regard this map of the plane as acting on the cylinder, or the torus, or finally on the rectangle

$$0 \leqslant x \leqslant 1$$
$$0 \leqslant p \leqslant 1/2 \qquad (5.3)$$

The rectangle, with edges identified appropriately, is topologically a sphere with singularities at four corners. We

will usually consider the standard map as acting on this manifold.

Another set of coordinates, useful for the computations, are the symmetry coordinates (x,y) [18]

$$y = p-(k/4\pi)\sin(2\pi x) \qquad (5.4)$$

which will be used for some of the plots. In coordinates (x,y) the standard map has perpendicular reflection symmetry about x=0.

When k=0 the standard map is integrable and the phase plane is foliated by invariant circles. As k increases an invariant circle with frequency ν will be destroyed at some critical value, $k_c(\nu)$, of the parameter. Circles with rational frequency, which typically have $k_c(m/n)=0$, are replaced by island chains, while circles with irrational frequency are replaced by cantori.

Numerical evidence [10,18-20] suggests that the most robust circles have noble frequency; that is $k_c(\nu)$ is locally maximum for ν noble. In particular, for the standard map in the phase rectangle the last rotational circle to be destroyed appears to be the "golden" circle with

$$\nu_g = 1/\gamma^2 \; ; \qquad \gamma = (1+\sqrt{5})/2 \; ; \qquad (5.5)$$

where γ is the golden mean. The calculated value of the critical parameter for this circle is [18]

$$k_g = k_c(\nu_g) \simeq 0.971635406 \qquad (5.6)$$

Twenty orbits of the standard map at k_g, including the golden circle, are shown in Fig.(5.1). In this section we will focus on

the golden cantorus which appears at $\Delta k = k-k_g > 0$. For Δk small we expect this cantorus to provide the most resistance to the transport from p = 0 to p = 1/2. The phase rectangle for a case with $\Delta k>0$ is shown in Fig.(1.2), where four orbits with initial conditions in the upper box are shown. These orbits lie in the connected irregular region, that is they extend from p=0.0 to p=0.5.

(b) Convergents

To compute ΔW and illustrate the turnstiles of the golden cantorus we obtain a sequence of periodic orbits with frequencies $m/n \to \nu_g$. The rationals m/n which minimise $|n\nu-m|$ over the set of all rationals with the same or smaller denominator, are given by the truncations of the continued fraction expansion: for $\nu = [a_0,a_1,...]$, the rationals

$$m_j/n_j = [a_0, a_1,...a_j] \qquad (5.7)$$

are called the convergents. The integer j will be referred to as the level of approximation. For ν_g the expansion is

$$\nu_g = [0,2,1,1,1,1,...] \qquad (5.8)$$

and the convergents are the ratios of two Fibonnaci series

$$m_j = 0,1,1,2,3,5,8,..$$

$$j = 0,1,2,3,...$$

$$n_j = 1,2,3,5,8,13,21,. \qquad (5.9)$$

In Fig (5.1) island chains corresponding to the first five convergents are visible.

(c) Symmetric Orbits

There are two techniques for finding orbits with a given frequency: one is to use the variational principle, Eq.3.13, and the other is merely to iterate the map, searching the phase plane for an initial condition which gives the correct frequency. We adopt the latter, and discuss the relative merits of each below.

Many maps, including the standard map, are reversible [18]. This symmetry can be exploited to limit the search for periodic orbits to one dimensional curves in the phase plane. A reversible map may be factored into the product of two orientation reversing involutions:

$$T = S_1 S_2 \; ; \qquad S_i^2 = I; \qquad \det(\nabla S) = -1. \quad (5.10)$$

The fixed points of the involutions form curves called symmetry lines. For the standard map the symmetry lines of interest are

$$S_1: \; x=0, \; x=1/2; \qquad S_2: \; x=p/2, \; x=p/2+1/2; \qquad (5.11)$$

Equation (5.10) implies that $S_i T S_i = T^{-1}$, so that the reflection of any orbit by S_i is the time reversal of another orbit of T. If an orbit is its own reflection it is called symmetric. Note, for example, that rotational invariant curves of the standard map (5.1) must be symmetric since they have points on the fixed lines of S_i. Similarly a symmetric periodic orbit must have a point on

some symmetry line [18]. Aubry [3] has shown that minimising orbits of a symmetric twist map are symmetric.

It is a remarkable observation [10] that for many maps there appears to be one particular symmetry line, called the dominant line, on which there is a point of a minimax orbit for every rational frequency. For the standard map (with $k \geqslant 0$) the dominant line is $x=0$. This is illustrated in Fig.5.1 by the islands, which surround the minimax orbits, centered on the dominant line.

6.COMPUTATIONS

(a)Techniques

The numerical calculations required for the application of the theory are complicated by problems of stability and precision, so we provide sufficient detail to show how these problems can be overcome.

We use the secant method to find minimising and minimax symmetric orbits. A symmetric orbit with $\nu=m/n$ is defined by a zero of the function

$$f(p_o) = x_n - m - X(p_n) \qquad (6.1)$$

where $X(p)$ defines a symmetry line, e.g. Eq.(5.11), and $x_o = X(p_o)$. To find the minimax orbit we choose the dominant symmetry line. The minimising orbit has a point on the line $x = 1/2$ for n odd and on $x=p/2$ for n even [18].

If an orbit is not very unstable, e.g. has a residue $|R|<1$, then the secant technique easily converges with reasonable initial guesses. When $k>k_c$, however, the residue of a convergent orbit to a cantorus increases approximately exponentially with k and finding the orbit is increasingly difficult. Since the residue is a measure of the eigenvalues of the linearization of T^n, Eq.(3.12), it indicates the intrinsic loss of precision for iteration. Supposing that $R>>1$, then approximately $\log_{10}(R)$ significant figures are lost each period. If we require the position of an orbit to N significant figures, using arithmetic with P digits of precision, then the residue is restricted to $R<10^{P-N}$.

Actually the computations are further restricted by the proliferation of nearby orbits which are not minimising or minimax, but have the same frequency. These orbits are born by tangent bifurcation. In order to find the minimising or minimax orbits, it is best to begin at some $k<k_c$ where these orbits are relatively isolated and use extrapolation to obtain a guess for the secant method at larger values of k. We use 6th order Lagrangian extrapolation and choose the step size in k so that the residue increases by roughly 20% each step (for example with $m/n=233/610$ a step $\Delta k= 3.125(10)^{-4}$ was used). The difficulty in following orbits restricts the residue in practice to $R\leqslant 10^{(P-N)/2}$.

Symmetry can be further utilized to achieve larger residue orbits. For the standard map there are four symmetry lines, Eq.5.11, and a symmetric periodic orbit must have a point on two of them. Definite rules for which lines contain which orbits have been obtained [18], which show that the two points are separated by roughly $n/2$ iterations. Therefore, using the initial and final symmetry lines gives an effective instability $\sqrt{R}$ and we can obtain orbits with $R\simeq 10^{P-N}$.

(b) Convergence

Convergence of the periodic orbits to a noble cantorus proceeds faster than geometrically. Letting a_j represent a property for the jth convergent to the cantorus (for example the position on the symmetry line, or ΔW) we observe

$$\alpha_j \simeq \alpha_\infty + b(R_*)^{-\gamma^j} \qquad (6.2)$$

where $R_*(k)$ is the same for all properties. The convergence exponent appears to be the golden ratio to at least 4 significant figures. The fact that Eq.(6.2) holds for ΔW corrects an assertion in [18] to the contrary.

The residues of the convergents, on the other hand, increase without bound for $k>k_c$ according to

$$R^{\pm}_j \simeq R^{\pm}(R_*)^{\gamma^j}, \qquad (6.3)$$

a result obtained by Greene [10], who calls R_* the mean residue. R_* is an intrinsic property of the cantorus and measures the average instability; its logarithm is the Lyapunov exponent. The residues of the minimising and minimax periodic orbits appear to have a universal ratio as $j\to\infty$

$$R^-/R^+ = -1.434 \qquad (6.4)$$

which can be regarded in some sense as the ratio of the residue of the cantorus to the homoclinic orbit.

Optimal results for a particular cantorus are obtained by careful choice of the precision N. Equations (6.2) and (6.3) together with b) $R^{\pm}\simeq O(1)$ implies that at level j we know the position of the cantorus to $O(1/R_j)$.

Since we are restricted to $R\leqslant 10^{P-N}$ and the error in the position of the jth convergent is 10^{-N}, the maximum precision in position of the cantorus is obtained by setting $N\simeq P/2$. In our case we use double precision arithmetic with P=28 and obtain the

position of the cantorus to about 14 significant figures.

(c)Results

An example of the computations for the standard map is given in Table (6.1), for $\Delta k=0.01$. Here we display the convergents to level 16 at which point the residue is $\simeq 10^{12}$. The third column shows the positions of the minimax orbits on the dominant line and gives, at level 16, the position of the cantorus to 15 places. The logarithm of the residues of both orbits, in column 4, are seen to increase in an approximate Fibonacci sequence as implied by Eq.6.3. The fifth column shows that the action differences decrease monotonically with level as expected from Fig(4.4). ΔW for the final level is omitted beause this orbit is too unstable to yield a reliable result: since ΔW is so small it is quite sensitive to orbit deviations. The final column gives the half width of the largest gap, which is centered on the dominant line. This was computed by looking for the point on the minimising orbit which falls nearest $x=0$.

The flux through the cantorus is most easily obtained by computing ΔW, but we can also construct the turnstiles to verify their shape. To obtain the turnstile for the largest gap of the level j orbit we join the endpoints of an iterate of the largest gap with a straight line. For an orbit of odd or even period we use the $((n-1)/2)^{th}$ or $(n/2)^{th}$ iterate of the major gap end points, and form the line from 100 points. Iterating these points both forward and backward to the major gap gives the curves C^- and C^+ respectively. The right half of the turnstiles for levels 6 through 10 is shown in Fig(6.1). Here we see that C^- and C^+ cross at the position of the minimax orbit, indicated by

a triangle in the figure. The convergence appears geometric in Fig(6.1) because the residues of levels $j \leqslant 9$ are still small: Note also that significant overlap of the turnstiles occurs beyond R=1.

The converged turnstile is shown in Fig(6.2) which is computed at j=15. Also shown are the iterates $T^t(C^+)$, $t>0$, and $T^t(C^-)$, $t<0$, which form the barrier. An important fact is that the turnstile is thin compared to the width of the gap: note the factor of 10^3 difference in scale between x and y in Fig(6.2). This implies that the flux through the cantorus is quite small, even though the gaps may be wide. This is because the motion is nearly parallel to the gap. Note also that the barrier does not appear to be a smooth curve, though C^+ and C^- are smooth. This might be expected since at k_c an invariant circle loses its smoothness [10].

Half turnstiles for $\Delta k=0.03$ and 0.05 are shown in Fig(6.3). Here the levels were j=13 and 12 respectively. Note that the shape of the cantorus in the neighbourhood of the endpoint of the major gap can be made identical in the two cases by a scale change; however, the shapes of the turnstiles appear to be different.

(d) Variational Methods

Finally we compare this technique for finding unstable periodic orbits with the use of the variational principle. Peyrard and Aubry [21] used the "gradient" technique to recast the variational principle for a periodic orbit of length n into a set of n differential equations in fictitious time which have the minimising periodic orbit as a fixed point. To solve for the

minimising orbit it is necessary to choose an initial condition known to be in the basin of this fixed point. The technique has the advantage that arbitrarily large residue orbits can be obtained. It has the disadvantage that solving n differential equations is clearly more work than iterating the map n times. Taking $n=377$ as an effective upper limit we see that the differential equation technique is more accurate than ours for $\Delta k > 0.05$. Of course as Δk increases the flux becomes larger and the cantorus less of a limiting factor in transport. Finding the minimax orbits by the gradient technique might appear to be harder, but if one knows that it has a point on some symmetry line, one can still minimise subject to that constraint.

7.MARKOV OR CIRCUIT MODEL FOR TRANSPORT

We have discussed a single cantorus, and the flux through it. Globally there are uncountably many cantori at all levels of the hierarchy. We would like to be able to formulate the global transport in terms of all of these cantori, but can not at present do so.

Our calculations indicate, however, that there are many problems in which the important transport properties are dominated by a discrete set of noble cantori. This leads us to an approximate picture that is radically different from previous pictures. In contrast to the usual continuous diffusion, we have a lumped element model, with regions of rapid diffusion separated by resistive barriers. In many applications, such as confinement in tokamaks and rates of chemical reactions, the internal diffusion in a region can be assumed to be infinitely rapid by comparison with the transport across the barriers. The validity of this model will be considered in section 10, with preliminary ideas on the treatment of the full problem.

We begin with general considerations about irregular components. The invariant circles of a given class separate the phase space into what Birkhoff called zones of instability, containing no circles of that class. Given one point on each boundary circle and neighbourhoods of each, he showed that there is an orbit which goes from one neighbourhood to the other. But a zone of instability does not typically have a dense orbit, as there can be islands in it. Removing all the islands leaves what we call an irregular component. Hopefully it has positive area, though this has not been proved. In practice one would neglect sufficiently small islands.

In seeking to understand the transport, we can treat the irregular components separately from each other, as there is no communication between them. To what extent is the motion in an irregular component chaotic? One definition of chaos for area preserving maps is existence of a Markov partition, i.e. a partition of the irregular component into regions with a list of allowable transitions between them, such that for each allowable sequence of regions there is a unique orbit, and the transition probabilities are independent of the past and future. Pesin [22] has shown that a Markov partition exists for components on which the Lyapunov exponents are non-zero almost everywhere. Numerical experiments do indicate nonzero Lyapunov exponents, though this has not been proved for any realistic examples. The significance of our present work is that it enables one to produce approximate Markov partitions, governing the long time transport, assuming small scale mixing.

The long time transport is limited by the strongest barriers. In fact the smallest ΔW on the route from A to B provides a strict upper bound on the transport. We propose that to estimate the transport it may suffice to consider a discrete set of barriers only, namely the most important noble cantori. This gives a lumped circuit model with regions R_i of area A_i connected by turnstiles. At each iteration the turnstiles push an area B_{ij} from region R_i to R_j, and an area B_{ji} the other way. The B_{ij} are given by the ΔW for the barrier, if we neglect the tiny islands within the turnstiles. For definiteness, the exit half of the turnstiles from region R_i should be included in A_i, but not the entrance. For zero nett flux:

$$B_{ij} = B_{ji} \qquad (7.1)$$

but we will not always need to assume this "detailed balance" condition. Area preservation, however, implies a general balance condition for each region R_i of finite area:

$$\sum_j B_{ij} = \sum_j B_{ji} \tag{7.2}$$

which we will always use.

Now besides the rotational cantori, there are vibrational cantori, which surround islands. If the island has period n, then these are invariant under T^n. Since the images are connected directly by cyclic permutation, we need represent them only once in the circuit model. The area to use for A is the combined area. The area to use for the turnstile is ΔW for one of them, because that is the area transported across one cantorus in n iterations, therefore equal to the area transported across all n in one iteration. Each island typically has its own islands round it, so one has an infinite hierarchy of regions connected by turnstiles in a tree, which is the topological dual of the barriers.

Let us assume that there is immediate loss of memory within the regions. Thus while in region i there is a chance:

$$p_{ij} = B_{ij}/A_i \tag{7.3}$$

at every step of landing in the turnstile that will take you to region j. Then the dynamics is that of a Markov chain. If the transition rates are small, as is probably necessary to justify the assumption of loss of memory, then one could view this as a continuous time Markov process. An isomorphic picture, which may be more familiar to some, is of an RC circuit. Suppose we have a charge distribution with charge Q_i in region i, or charge density:

$$V_i = Q_i/A_i \tag{7.4}$$

Then the flux from i to j is V_iB_{ij}, and from j to i is V_jB_{ji}. Thus:

$$\frac{dQ_i}{dt} = \sum_j V_jB_{ji} - V_iB_{ij} \tag{7.5}$$

$$= \sum_j (V_j - V_i)B_{ji} \tag{7.6}$$

from equation (7.2). So this looks just like a bunch of capacitors of capacitance A_i, with their bottom plates all connected to ground, and their top plates connected by resistors of conductance B_{ij} or resistance $1/B_{ij}$.

8. RELAXATION TO EQUILIBRIUM

What predictions can such a Markov model make? Firstly, it predicts relaxation to equilibrium with V_i constant, Q_i proportional to A_i, if the total area is finite. It can be useful to consider irregular components with some regions of infinite area, e.g. representing escape from a tokamak, or a completed chemical reaction. We will restrict attention for simplicity to systems with finite regions, apart from the possible exception of one region of infinite area.

If the total area is finite, then relaxation to equilibrium predicts ergodicity: the fraction of time spent in any region is proportional to its area. Furthermore, it predicts mixing: any initial charge distribution converges to A_i.

Next we discuss the rate of relaxation to equilibrium. Generally, the relaxation to equilibrium is a sum of exponential decays, the rates being the logarithms of the eigenvalues of the transition matrix p (7.3), in discrete time, and the eigenvalues of p-I in continuous time (I being the identity matrix), or equivalently, the inverse of the RC time constants of the circuit. The rates are of the order of B_{ij}/A_i. If p has non-trivial Jordan blocks, they give polynomial times exponential decays, and if there are infinitely many states and p-I is not compact, then one can get other forms of decay.

To solve equations (7.6) take the Laplace transform. In discrete time, take the generating function:

$$P_{ij}(s) = \sum_{t \geq 0} (p^t)_{ij} s^t \qquad (8.1)$$

for the t-step transition matrix p^t. Conditioning on the first step gives, in matrix notation:

$$P(s) = I + s\, p.P(s) \qquad (8.2)$$

which can be solved for P(s). In either case, the singularities of the transform or generating function give the decay, e.g. simple poles give exponential decay.

For irregular components bounded by invariant circles, there appear to be arbitrarily large time constants. This is because near an invariant circle there are cantori with arbitrarily small ΔW, by continuity, and apparently the ΔW tend to zero faster than the areas of the regions between them. For example, if there is a critical noble circle on the boundary, self-similarity [29] implies that there will be a sequences of time constants scaling like γ^{2i}, where i labels the barriers. This will be further discussed in section 11.

The boundary circles must be critical circles (i.e. non-smooth and destructible by arbitrarily small perturbation), because smooth circles have others arbitrarily close on both sides. But they need not all have noble rotation number. Nevertheless, we expect some kind of universal self-similarity [23].

We propose that this is the origin of the observed long time tails to the decay of correlations [9,15,24,25]. The "elementary" correlation functions are precisely the elements of p^t, and the correlation between the distribution evolved after time t from a distribution h_i, with a distribution g_i is just $h^T p^t g$ (superscript T meaning transpose). Usually one chooses distributions with zero nett weight. For some applications, only a range of timescales may be important. Then if a barrier between regions i and j has a turnstile of area ΔW_{ij} and $A_i/\Delta W_{ij}$,

$A_j/\Delta W_{ij}$ are both large by comparison with the maximum time of interest, the barrier can be regarded as inpenetrable, thus truncating the distribution of time constants. Rather than truncation, however, it may be better to approximate the neighbourhood of the boundary circles by some infinite self-similar hierarchy.

Most of the observations of decay of correlations have been interpreted as power law decays. One might ask how a sum of exponentials can look like a power law, but is just a question of the size of the coefficients. The formula:

$$\int_0^\infty d\tau\ \tau^{-(1+\mu)} \exp(-\frac{t}{\tau}) = \Gamma(\mu)\ t^{-\mu} \qquad (8.3)$$

is a familiar illustration. In our case we are considering a discrete set of time constants, but they can still give power laws e.g. [26]:

$$\sum_{n=-\infty}^{\infty} a^{-n} \exp(-b^n t) = t^{-\log_b a}\ Z(\log_b t) \qquad (8.4)$$

where Z is some periodic function of period 1.

We are not yet able to predict what power laws one should expect for the decay of interesting correlations. This would require an understanding of the self-similarity near general boundary circles, and of the relative importance of the different classes of boundary circle.

It is worth mentioning here that a lot of the work that has been done on correlation functions $C(\tau)$ has been to evaluate a "diffusion constant", as:

$$D = \int_0^\infty d\tau\ C(\tau) \qquad (8.5)$$

In situations where the diffusion is so inhomogeneous, there may not be much value in such an average diffusion constant. The flux

across the barriers is already given by a diffusion or conductivity relation:

$$\text{flux} = \text{conductivity} \times \text{difference in densities} \qquad (8.6)$$

and we derive the decay of correlations from this, rather than the other way round.

9. TRANSIT TIMES

Another very useful prediction that the Markov model can make concerns the time orbits take to get from one part of an irregular component to another. We define the first visit time from i to j to be the first time of visiting j starting in i. The first return time to a state i starting in i has various definitions depending whether you want just the next time in i (makes sense only in discrete time), or the first time in i having been somewhere else, or the duration of the first interval not in i. We will discuss the distribution of these transit times, and their expectations.

First visit times are particularly important for systems with an escape region. Of course, the distribution of escape times can be derived from the rate of relaxation of the charge in the escape region, as discussed in section 8. But it is worth a separate discussion. In general, one can find the distribution of first visit times to a state j by considering a modified chain in which j is made absorbing (infinite area). One finds a sum of exponential decays again, though not necessarily the same ones, unless j was really absorbing.

Coming now to the first return times to state i, their distribution can be computed fairly obviously from the transition probabilities to its neighbours, and a knowledge of the first visit times from the neighbours to i.

There are interesting results for the expectations of the first visit and first return times, which are particularly simple. Start with first return times to i. Let:

$$A = \sum_j A_j \qquad (9.1)$$

be the total area accessible to i, and:

$$B_i = \sum_j B_{ji} \tag{9.2}$$

be the total area of the turnstiles into i. Then the expected next time in i, starting in i (in discrete time) is:

$$A/A_i \tag{9.3}$$

The expected first time in i starting in i, having been elsewhere is, in both discrete and continuous time:

$$A/B_i \tag{9.4}$$

and the expected duration of the first interval not in i is:

$$(A-A_i)/B_i \tag{9.5}$$

So remarkably, the only dependence on the regions outside i is through their total area. The details of connections and time constants etc. is irrelevant to expected return times. Of course this shows that expected return times need to be used with caution, as physically it wouldn't make any difference if the accessible area were increased via a very slow turnstile, say. These results are a restatement in our terms of standard results of Markov theory, e.g. [27]. They require only the balance assumption (7.2).

Similarly, provided $B_{ij} = B_{ji}$, and all routes from i to j have to pass through the turnstile from i to j, the expected time from a state i to a neighbour j has a very simple form. To calculate it, remove the turnstile. This reduces A_i by B_{ij}, and

leaves i in a component satisfying the balance assumption, so in discrete time the expected next time in i is:

$$E_{ij} = \frac{C_{ij} - B_{ij}}{A_i - B_{ij}} \qquad (9.6)$$

where C_{ij} is the area accessible to i without entering j (but including B_{ij}). Now each time the system is in state i it really has a chance B_{ij}/A_i of hopping into j. So condition on this and its alternative to get the following equation for the expected time t_{ij} from i to j:

$$t_{ij} = \frac{B_{ij}}{A_i}.1 + (1 - \frac{B_{ij}}{A_i})(E_{ij} + t_{ij}) \qquad (9.7)$$

Solving for t_{ij} gives:

$$t_{ij} = C_{ij}/B_{ij} \qquad (9.8)$$

another remarkably simple and useful result. The same result holds in continuous time.

To get the expected first time between states which are not neighbours, simply add up the expected first times for each transition along the direct route between them.

As an example, figure 1.2 shows the result of feeding particles in at the neighbourhood of the unstable two-cycle in the standard map and removing them when they get near the unstable fixed point. From the sudden changes in density it is clear that there are just a few major barriers, with relatively rapid transport in between. It is like a system of beaver dams. The dams correspond to noble cantori, as illustrated. Actually, it is the minimax quasiperiodic orbits that were plotted, rather than the cantori, because it turned out to be easier to get more points

on them. They were located by a bisection method on the line $x=0.0$, looking for a point whose n_j^{th} iterates lay on the right or left of $x=m_j$, according as the rational approximant m_j/n_j was greater or smaller than the frequency of interest.

The average length of 27 orbits from the upper box to the lower one was around 6000. If one wants the expected time from the two-cycle to the fixed point, the above results show that one need worry only about rotational barriers and the accessible area above them. We need not consider any vibrational cantori. Using the techniques of section 6, we measured ΔW for the golden cantorus to be 5.78×10^{-5}. By subdividing figure 1.1 into little squares, we estimated the accessible area above the golden cantorus to be about 0.063 (remember the phase rectangle has an area of 0.5). Thus we expect the time to cross from just above the golden cantorus to just below to be:

$$0.063/(5.78\times10^{-5}) \approx 1090 \qquad (9.9)$$

iterations. Clearly, the golden cantorus alone is insufficient to explain the observed transit time.

If we include some more cantori, as in table 9.1, we get a total expected transit time of about 9500, which is closer, though now an overestimate. This could be due to overestimating the areas A, because there may be a significant density of small holes. In table 9.1, the rotation numbers of the cantori are nobles, and are specified by the first two convergents m_J/n_J, m_{J+1}/n_{J+1} such that $a_j=1$ for $j\geq J+2$. We call them the initial convergents for the noble. The cantori are not shown in figure 1.1, but knowing the initial convergents you can work out where they are relative to the island chains. Table 9.1 also includes

an estimate of the time to cross the edges of the lower box. It has an area of 0.005, so we supposed the turnstile for entry has area around 0.004.

10. MODIFICATIONS TO THE MODEL

Our lumped circuit model appears to give a reasonable first approximation. The discretisation procedure is somewhat unsatisfactory, however, because as one includes the effect of more cantori, the expected transit time increases without limit. Indeed, the result of table 9.1 already overestimates the transit time. The resolution is that when the barriers are close, a turnstile may take you across more than one barrier at once. This gives a more complicated circuit, with resistors joining regions that are not neighbours. We have not fully analysed these complications, but we sketch some preliminary ideas.

From the practical point of view, one would like to obtain a cutoff condition, telling one how many of the cantori one should keep to get the same effective transport.

As for the basic theory, we outline two approaches to the treatment of the complete circuit, with all the cantori.

In the first approach, the turnstiles are chosen to be aligned in "chimneys". The motion outside the chimneys looks like rotation on invariant circles. After a number of iterations, however, the orbit arrives at a chimney, where the upward and downward motion constitute a Markov process.

In the second approach, one considers a more general form of partial barrier through cantori, for which the flux is distributed over all the gaps. Any choice of partial barriers defines an action coordinate by the area underneath it. The best choice of partial barriers, from the point of view of longtime prediction, would be one which makes successive steps in the action as independent as possible. If one assumes full independence of the steps, this would give a Markov process in

continuous action space. But how to choose the partial barriers, and to find the distribution of the steps in action is unknown at present.

In both approaches, inclusion of cantori of all classes in the hierarchy is no problem. The Markov process simply takes place along the branches of a tree.

One might ask whether the set of partial barriers constructed from cantori of all classes is dense. Or are there regions with no cantori? Whenever there is twist about some point, there are circles or cantori. For example, around typical elliptic periodic points. But hyperbolic periodic orbits have no neighbourhood in which the map twists around them, because their invariant manifolds do not twist. Thus, when islands disintegrate as their residue increases through +1, a region develops containing no cantori. However, it typically contains a horseshoe, which is effective at mixing up the orbits, which is just what we want in the regions between barriers.

11. SCALING LAW NEAR CRITICAL

If there is an invariant circle, then it separates the regions on either side. As parameters are varied, however, the circle may break into a cantorus, permitting flux from one side to the other. Apparently, the last circle to break between any two regions always has noble rotation number. The breakup of noble circles appears to be governed by a universal one parameter family, exhibiting universal self-similarity and scaling behaviour [29]. In particular this leads to a scaling law for the flux through a cantorus, near criticality.

Suppose k_c is a critical parameter value for a noble frequency ν. Then there appear to exist scaling factors:

$$\begin{aligned} \alpha &\simeq -1.4148360 \\ \beta &\simeq -3.0668882 \\ \delta &\simeq 1.6279500 \end{aligned} \qquad (11.1)$$

such that for Δk small:

$$\Delta W_\nu(k_c+\Delta k/\delta) \simeq \Delta W_\nu(k_c+\Delta k)/\alpha\beta \qquad (11.2)$$

In fact, there appears to be a universal periodic function:

$$U(x) = U(x+1) \qquad (11.3)$$

such that there exist scales A in area and Δk_0 in parameter such that:

$$\Delta W_\nu(k_c+\Delta k) \simeq A\ (\Delta k/\Delta k_0)^\eta\ U(\log_\delta(\Delta k/\Delta k_0)) \qquad (11.4)$$

where $\eta = \log_\delta \alpha\beta \simeq 3.0117220$ (11.5)

The universal function U is very close to constant [18]. Thus for practical purposes U can be treated as constant, and it is not necessary to find Δk_0.

This scaling law for ΔW leads to important conclusions for the flux. In the simplest case when we ignore all other cantori in the neighbourhood, this gives a conductivity scaling like Δk^η. Even if we keep all the cantori in the neighbourhood, the scaling properties of the universal one parameter family imply that for every resistor at parameter $k_c+\Delta k$ there is one $\alpha\beta$ times larger at $k_c+\Delta k/\delta$. Of course, a new one comes in at the outside, but if the total effect out to infinity is convergent, one could regard oneself as being at infinity. The areas of the regions between them also scale, but since in the formula for expected time from one region to another all the numerators are dominated by the area of the initial region, this gives an expected transit time scaling like $\Delta k^{-\eta}$.

Since η is fairly large, this means that the flux grows very slowly with k. This explains why it is hard to determine critical parameter values for circles by looking for orbits to cross from one side to the other [30]. Chirikov [9] measured average transit times from $p\simeq 0$ to $p\simeq 0.5$ in the standard map (considered on the torus). He fitted his results to a scaling law of the form:

$$\text{Expected time} \simeq C\ (k-k_c)^{\mu} \qquad (11.6)$$

the best fit giving $k_c\simeq 0.989$, $\mu\simeq 2.55$. For small k, however, his

results are also consistent with exponent η and critical value:

$$k_c \simeq 0.972 \qquad (11.7)$$

corresponding to Greene's conjecture [10] that the last circle to break is the golden one.

In figure 11.1, we see that apart from statistical fluctuations, the flux given by our scaling law with a constant determined by the turnstile of the golden cantorus is an effective lower bound on the numerical results of Chirikov. For k far enough from k_c, however, deviations from the scaling law become apparent.

How far from critical the scaling law will be a good approximation depends on the particular system. For the standard map and rotation number $1/\gamma^2$, we find that ΔW differs from the scaling result by 10% for $\Delta k = 0.1$, and 20% for $\Delta k = 0.15$. The best we can do in general is to present the form of the dominant corrections to the scaling law for ΔW. These are of two forms:

(i) Intrinsic corrections: A given system is described by the universal one parameter family only close enough to critical. The dominant deviations are given by a ratio [18]:

$$\delta' \simeq -0.61083028 \qquad (11.8)$$

For example, $$\Delta W(k_c+\Delta k/\delta^n) \simeq A/(\alpha\beta)^n \ (1 + \delta'^n) \qquad (11.9)$$

Thus there are correction terms to $\Delta W(k_c+\Delta k)$ of the form:

$$C\ \Delta k^{\log_\delta \alpha\beta/|\delta'|}\ U_2(\log_\delta \Delta k) \qquad (11.10)$$

where U_2 is a universal function satisfying $U_2(x+1)=-U_2(x)$. The exponent is $\eta+1.0115...$ There is nothing one can say in general about the coefficient C of the correction.

(ii) Removable corrections: If one had a system satisfying the scaling law, then a general reparametrisation and/or change of scale with parameter would introduce corrections of the form $\Delta k^{\eta+1}$ and higher order. Thus if one were to determine the coefficients of such terms one could remove them by reparametrisation and/or parameter dependent scale change.

Note that although the exponents for the two types of correction are extremely close, the first correction has an oscillatory behaviour whereas the second does not, so they are distinguishable.

12. DISCUSSION

Cantori, partial barriers, turnstiles and chimneys, together with the Markov and circuit models, provide a picture and a theory of irregular motion that differ in many respects from others. They help to explain the complexity and also the surprising degree of order in the motion. They provide a means of estimating the widely differing rate constants for transport between different regions of phase space separated by cantori, and resolve the apparent discrepancy between different estimates of the critical value of the parameter for the golden circle of the standard map.

Preliminary calculations and comparison with the work of Chirikov show that a simple circuit model with series resistors representing successive cantori is enough to obtain the mean rate for transport between two regions. However, a single rate constant can be very misleading, as the mean can be made up from the widely differing rate constants of different types of process. But more detailed calculations are needed in order to obtain rate constants for these.

The formal extension of the theory and methods to Hamiltonian systems of two degrees of freedom is simple. There is no need to construct a map explicitly by using a surface of section. The orbits and action integrals can all be obtained by direct numerical integration, and all the limiting properties are the same, so the crucial quantities ΔW can be found. The variational principles and methods can also be extended in a natural way, e.g. [31].

This form of the theory can be used directly for plasma containment in Tokamaks, to analyse the so-called stochastic

component of a static magnetic field, and guiding centre motion. In order to estimate rates of heat or particle diffusion, it would be necessary to extend the model to take into account other effects. These would include collisions, neoclassical effects and adiabatic variations in the magnetic field.

Particles in accelerators, asteroids near Kirkwood gaps and atoms involved in chemical reactions have more than two degrees of freedom, but often essential features of the motion can be represented by a reduced system, so that our theory can be used.

We speculate on some possible future developments and questions raised.

The presence of partial barriers with small turnstiles changes the usual picture of the relation between Lyapunov exponents and transport. The Lyapunov exponent and the entropy are measures of local mixing, and primarily measure the rate of mixing within a given region bounded by cantori. But for many purposes this relatively rapid process is irrelevant. What matters in molecular reactions is the rate of transport between regions of phase space representing different molecules or the same molecule in a substantially different configuration. What is wanted in plasmas or accellerators is the rate of escape from large regions of phase space, representing motion within some substantial region of configuration space.

The most important rate constants are usually those for transport between these regions. These are determined by the ratio of areas of turnstiles to areas of regions, and have little to do with Lyapunov exponents, which primarily measure the rate of divergence within regions. Neighbouring orbits commonly take very little time to separate to distances similar to the size of a

region bounded by cantori, but very much longer to penetrate through a turnstile into another region. The Lyapunov exponents give the rate at which a nice cat in a region of phase space turns into a mixed up cat. The turnstile rate constants give the mean rate at which bits of the cat are transported to regions of the phase space on the other side of cantori.

How to extend the ideas of this paper to higher dimensions is an interesting problem. There are higher dimensional cantori, and the ΔWs probably generalise to a vector of areas. But the almost invariant tori one would construct can not separate the phase space, so their importance is unclear.

A major part of the significance of this paper is that it indicates how to construct an approximate action variable or adiabatic invariant, together with a measure of how fast it can change. This is a problem of long-standing interest, e.g. Prigogine, Nekhoroshev.

Reinhardt et al [32] point to the importance of approximate tori in the semiclassical theory of the irregular regions of phase space, and its relevance to the properties of molecules. We are able to refine their criterion for semiclassical quantization. The quantization is limited by the size of the turnstiles in units of Planck's constant. Since the turnstiles vary in a very complicated way as a function of frequency, this criterion is difficult to apply in practice. This same complication provides a possible explanation as to why it has been so difficult to find a satisfactory criterion for the validity of semiclassical mechanics for real nonlinear systems such as molecules.

ACKNOWLEDGEMENTS

This research was partially supported by the Science and Engineering Research Council of the U.K. and by the U.S.DOE contract DE-FG05-80ET-53088. We would like to thank L.Kadanoff for open discussion of his parallel work with Bensimon, J.Mather for useful conversations, C.Murray for suggesting the term "turnstile", and B.Chirikov and J.Ford for very constructive comments on the manuscript. JDM acknowledges the hospitality and support of Culham Laboratory through grant CUL-341.

REFERENCES

[1] I.C.Percival, in Nonlinear Dynamics and the Beam-Beam Interaction, M.Month and J.C.Herrera, eds., Am. Institute of Physics Conf. Proc. No. 57 (1979) 302.

[2] S.Aubry, in Solitons and condensed matter physics, eds A.R. Bishop & T.Schneider (Springer 1978) p.264.

[3] S. Aubry and P.Y. Le Daeron, Physica 8D (1983) 381

[4] J.N. Mather, Topology 21 (1982) 457

[5] A. Katok, Ergodic Theory & Dyn.Sys. 2 (1982) 185

[6] E. Wigner, J. Chem Phys 5 (1937) 720; see also R. Marcelin, Ann.Physique 3 (1915) 120.

[7] D.Bensimon and L.P. Kadanoff, "Extended chaos and disappearance of KAM trajectories", preprint, Chicago (1983)

[8] J. N. Mather, "A criterion for non-existence of invariant circles", preprint (1982), Princeton

[9] B. V. Chirikov, Phys Reports 52 (1979) 263.

[10] J. M. Greene, J. Math Phys. 20 (1979) 1183.

[11] J. C. Keck, Adv. Chem Phys. 13 (1967) 85.

[12] R. S. MacKay & J. D. Meiss, Phys. Lett. 98A (1983) 92.

[13] I. C. Percival, J. Phys 12A (1979) L57.

[14] G. D. Birkhoff, Trans AMS 18 (1917) 199.

[15] S. R. Channon & J. L. Lebowitz, NY Ac Sci 357 (1980) 108.

[16] J. H. Bartlett, Cel. Mech. 28 (1982) 295.

[17] Scholtz, private communication

[18] R. S. MacKay, "Renormalisation in area preserving maps", Ph.D. thesis (1982) Princeton (Univ. Microfilms Int., Ann Arbor, Michigan)

[19] G. Schmidt, Phys Rev. 22A (1980) 2849.

[20] I. C. Percival, Physica 6D (1982) 67.

[21] M. Peyrard & S. Aubry, "Critical behaviour at the transition by breaking of analyticity in the discrete Frenkel-Kontorova model", submitted to J Phys C.

[22] Ya. B. Pesin, Russ Math Surveys 32 (1977) 55.

[23] R. S. MacKay & I.C. Percival, "Universal scaling for the boundary of Siegel domains of arbitrary rotation number", to appear.

[24] J. D. Meiss, J. R. Cary, C. Grebogi, J. D. Crawford, A. N. Kaufman & H.D.I Abarbanel, Physica 6D (1983) 360.

[25] C.F.F. Karney, Physica 8D (1983) 360; F.Vivaldi, G.Casati and I.Guarneri, Phys.Rev.Lett. 51 (1983) 727

[26] J. M. Greene, private communication

[27] J. G. Kemeny, J. L. Snell & A. W. Knapp, Denumerable Markov Chains (Springer, 1976).

[28] A.B.Rechester & R.B.White, Phys.Rev.Lett 44 (1980) 1586.

[29] R. S. MacKay, Physica 7D (1983) 283.

[30] D. F. Escande & F. Doveil, J. Stat Phys 26 (1981) 257.

[31] R. H. G. Helleman, in Topics in Nonlinear dynamics, ed. S.Jorna, AIP conf proc 46 (1978) 264.

[32] W. P. Reinhardt, J. Phys Chem 86 (1982) 2158; R. B. Shirts & W. P. Reinhardt, J Chem Phys 77 (1982) 5204; C. Jaffe & W.P.Reinhardt, J. Chem. Phys 77 (1982) 5191.

TABLE CAPTIONS

Table 6.1: Various quantities for the periodic orbits with rotation numbers ν equal to convergents to $1/\gamma^2$, in the standard map at $k= k_g+0.01$. Tabulated are the position p of the minimax periodic point on the dominant line, x-coordinate x_g of the nearest minimising periodic point, residues $R_\pm$ of the two orbits, and their difference in action ΔW.

Table 9.1: Initial convergents to some nobles, ΔW for the corresponding cantorus, accessible area A above it, and consequent expected time t to cross from above to below, for the standard map at $k=1.121635$.

Table 9.1

initial	convergents	$\Delta W/10^{-4}$	A	t
1/2	4/9	1.71138	0.035	200
1/2	3/7	1.1891	0.04	340
1/2	2/5	0.711096	0.045	630
2/5	5/12	0.7066	0.05	710
2/5	5/13	0.625048	0.056	900
1/2	1/3	0.57838	0.063	1090
1/3	3/8	0.7966	0.075	940
1/3	4/11	1.0773	0.09	840
1/3	3/10	1.77204	0.105	590
1/3	2/7	1.576	0.117	740
0/1	1/3	1.759	0.135	770
1/4	3/11	2.0218	0.15	740
1/4	2/9	3.9086	0.17	430
0/1	1/4	4.262	0.185	430
lower box		40.0	0.274	70
total				9420

FIGURE CAPTIONS

Figure 1.1: One orbit of the standard map for k=1.121635

Figure 1.2: Some orbits of the standard map for k=1.121635. Four orbits were started in the upper box and stopped when they first entered the lower box. The large dots are orbits homoclinic to cantori of rotation numbers $1/\gamma^2$ and $(1+\gamma)/(4+3\gamma)$.

Figure 1.3: A cantorus of the standard map at $k=1.001635=k_g+0.03$ for $\nu=1/\gamma^2$

Figure 2.1: Illustration of the flux across a closed curve

Figure 2.2: Illustration of the flux across a family of closed curves

Figure 2.3: Partial barrier with turnstile

Figure 3.1: Geometrical interpretation of the generating function

Figure 4.1: Formation of a partial barrier with turnstile from periodic orbits

Figure 4.2: $\Delta W_{m/n}$ for a selection of rationals for the universal map for the neighbourhood of critical noble circles.

Figure 4.3: Formation of partial barrier with turnstile from a cantorus

Figure 4.4: Formation of partial barrier with turnstile from a broken separatrix

Figure 5.1: Twenty orbits of the standard map for $k=k_g$, including an orbit on the golden circle.

Figure 6.1: Half-turnstiles for approximant periodic orbits of levels 5 to 9 to the golden cantorus in the standard map at $k=k_g+0.01$

Figure 6.2: Partial barrier with turnstile constructed from the golden cantorus in the standard map at $k=k_g+0.01$

Figure 6.3: Half-turnstiles for the golden cantorus in the standard map at $k=k_g$ + a) 0.03; b) 0.05

Figure 11.1: Comparison of Chirikov's numerical experiments and our formulae. N is the number of iterations to cross the region near the golden cantorus and k is the parameter. The continuous line is for our scaling law (11.7) fitted to the value for the golden cantorus alone, giving $N=25(\Delta k)^{-\eta}$.

DOE/ET-53088-109

IFSR #109

TRANSPORT IN HAMILTONIAN SYSTEMS

R. S. MacKay[•], J. D. Meiss, and I. C. Percival[•]
Institute for Fusion Studies
University of Texas
Austin, Texas 78712

[•] Queen Mary College, London E1 4NS

September 1983

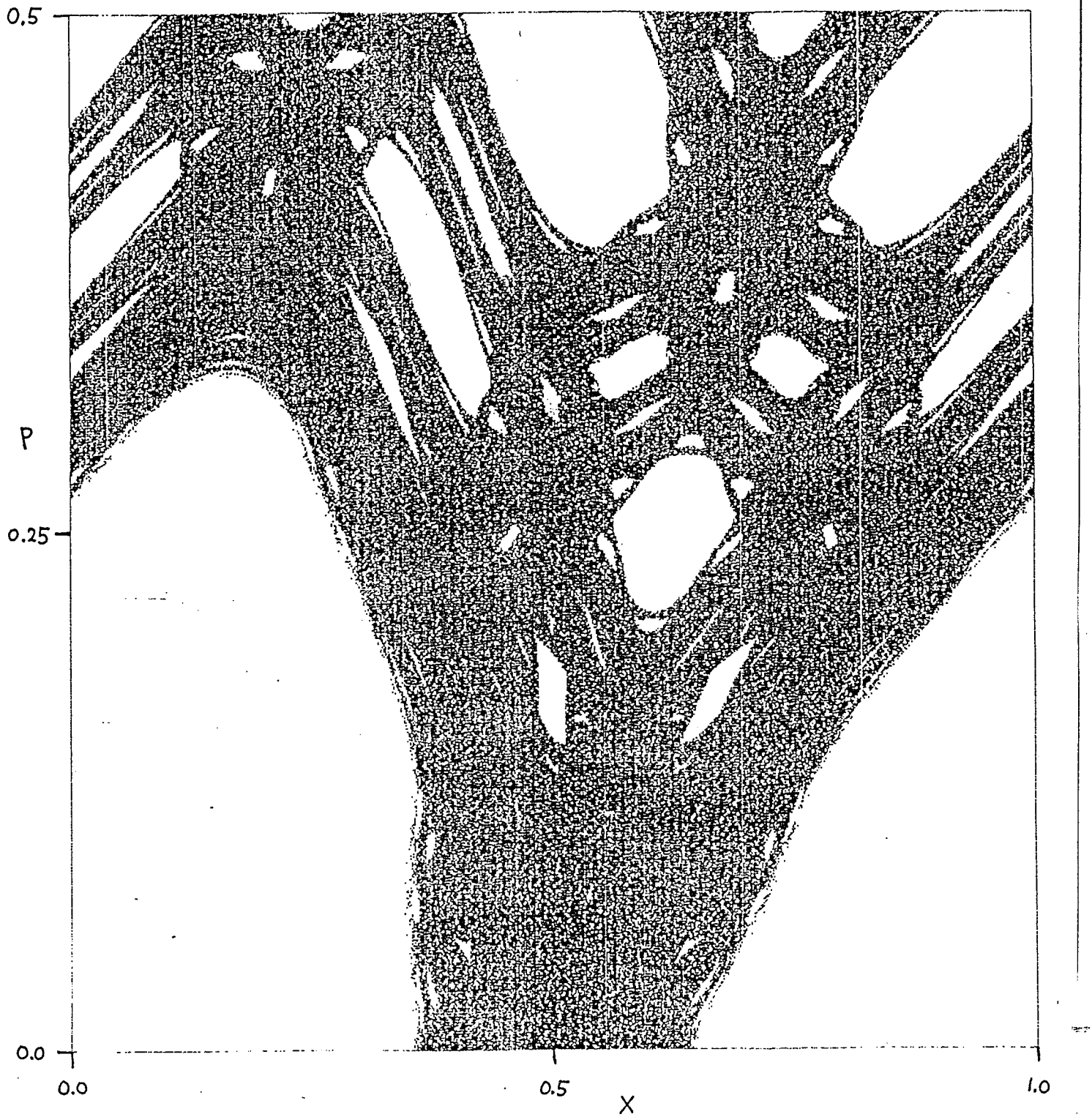

FIG. 1.1

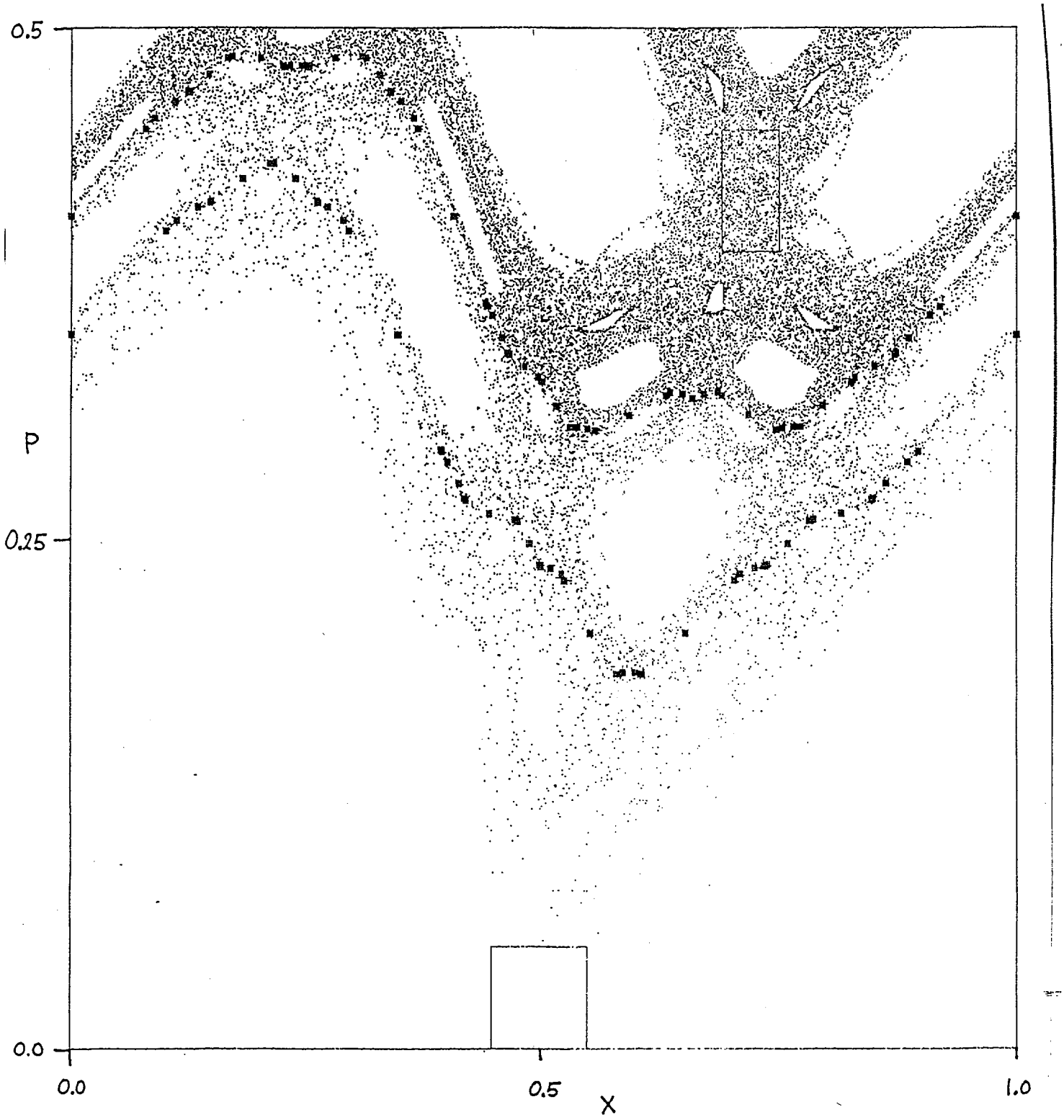

FIG. 1.2

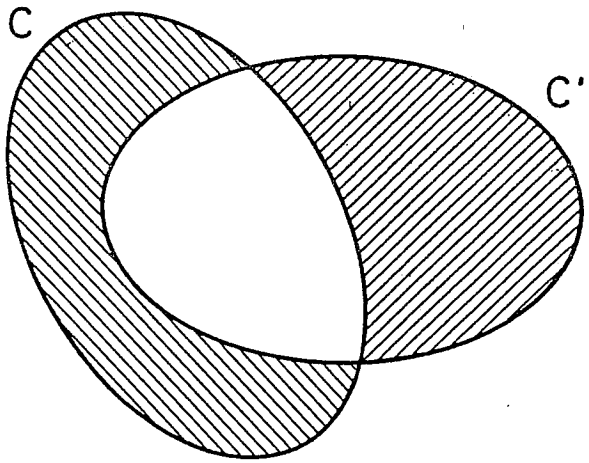

FIG. 2.1

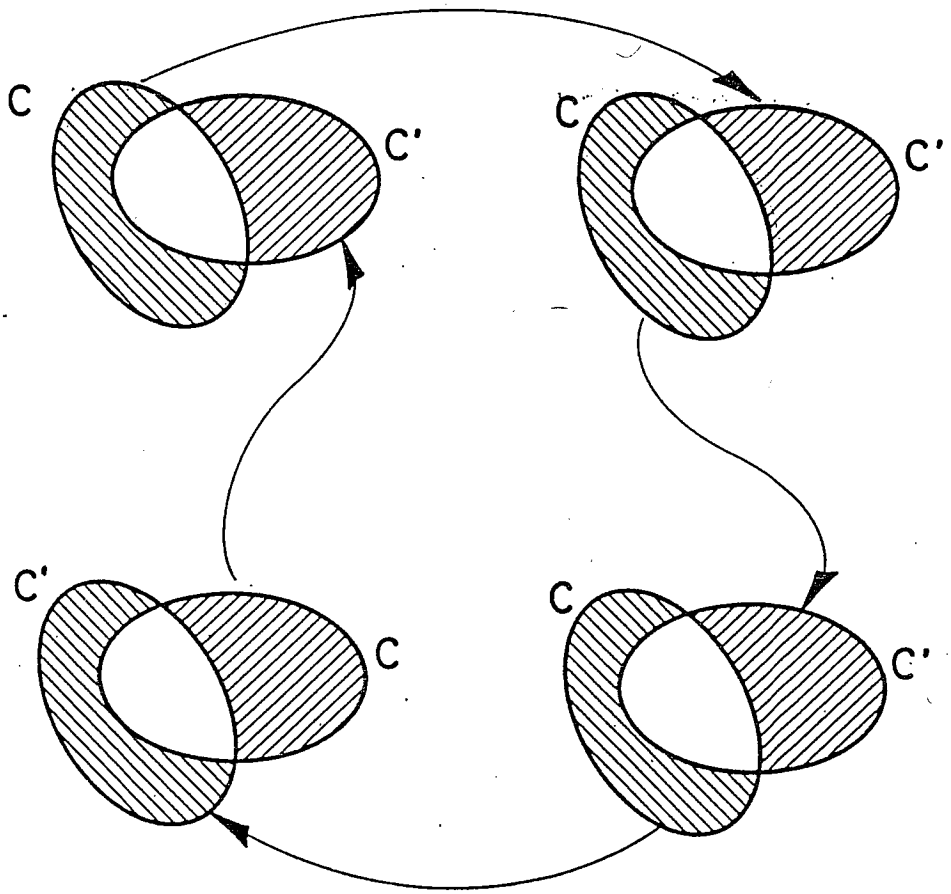

FIG. 2.2

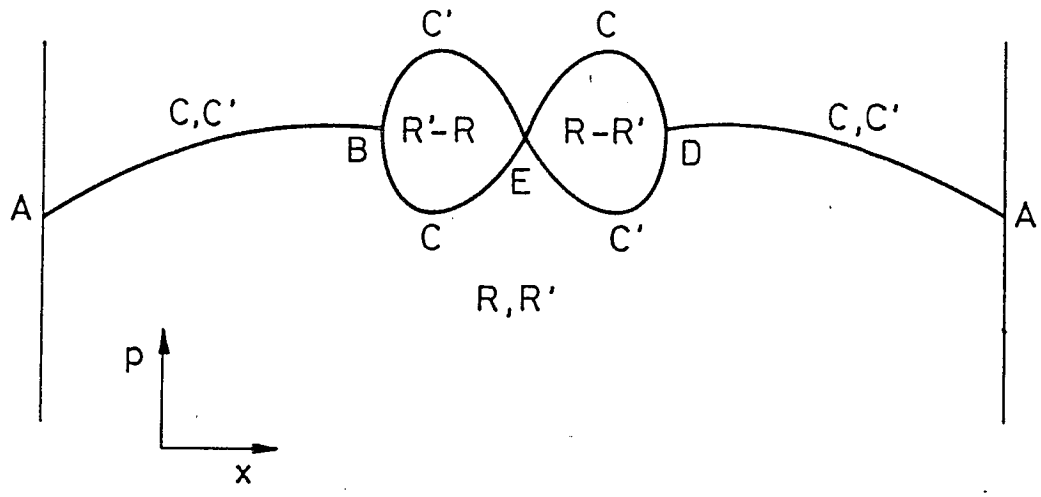

FIG. 2.3

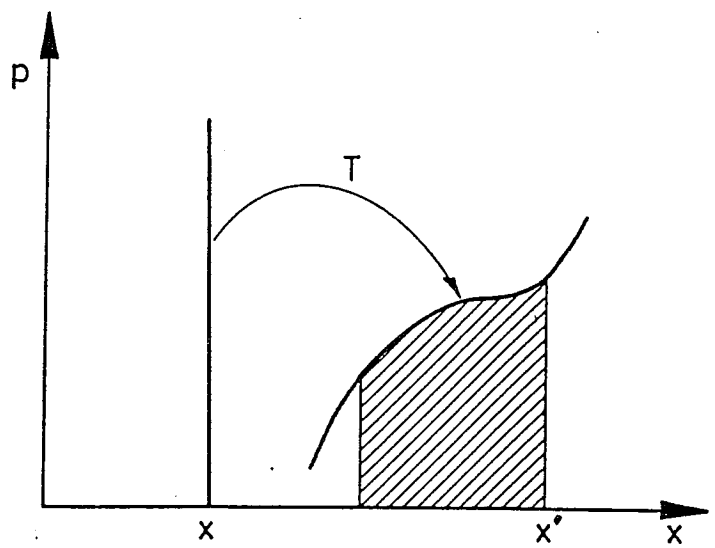

FIG. 3.1

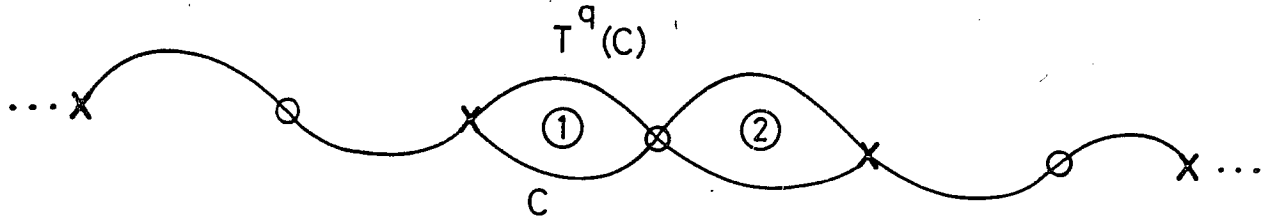

FIG. 4.1

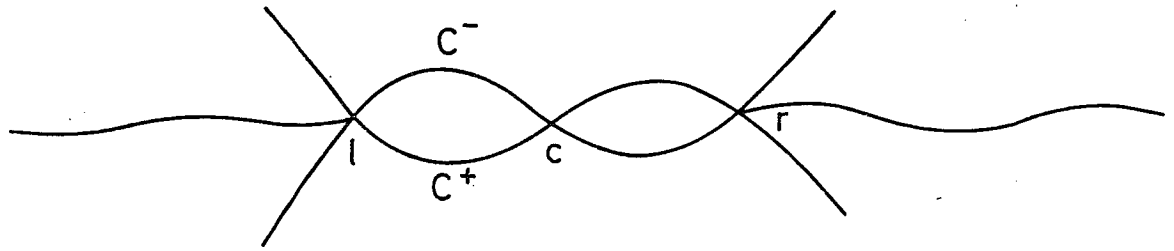

FIG. 4.3

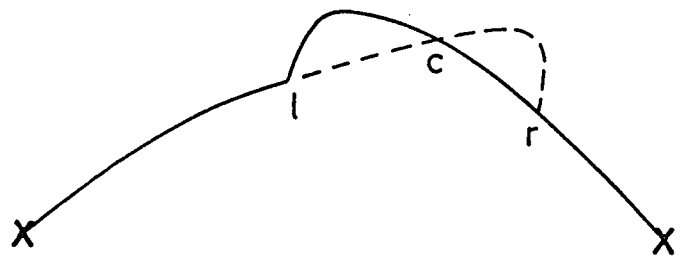

FIG. 4.4

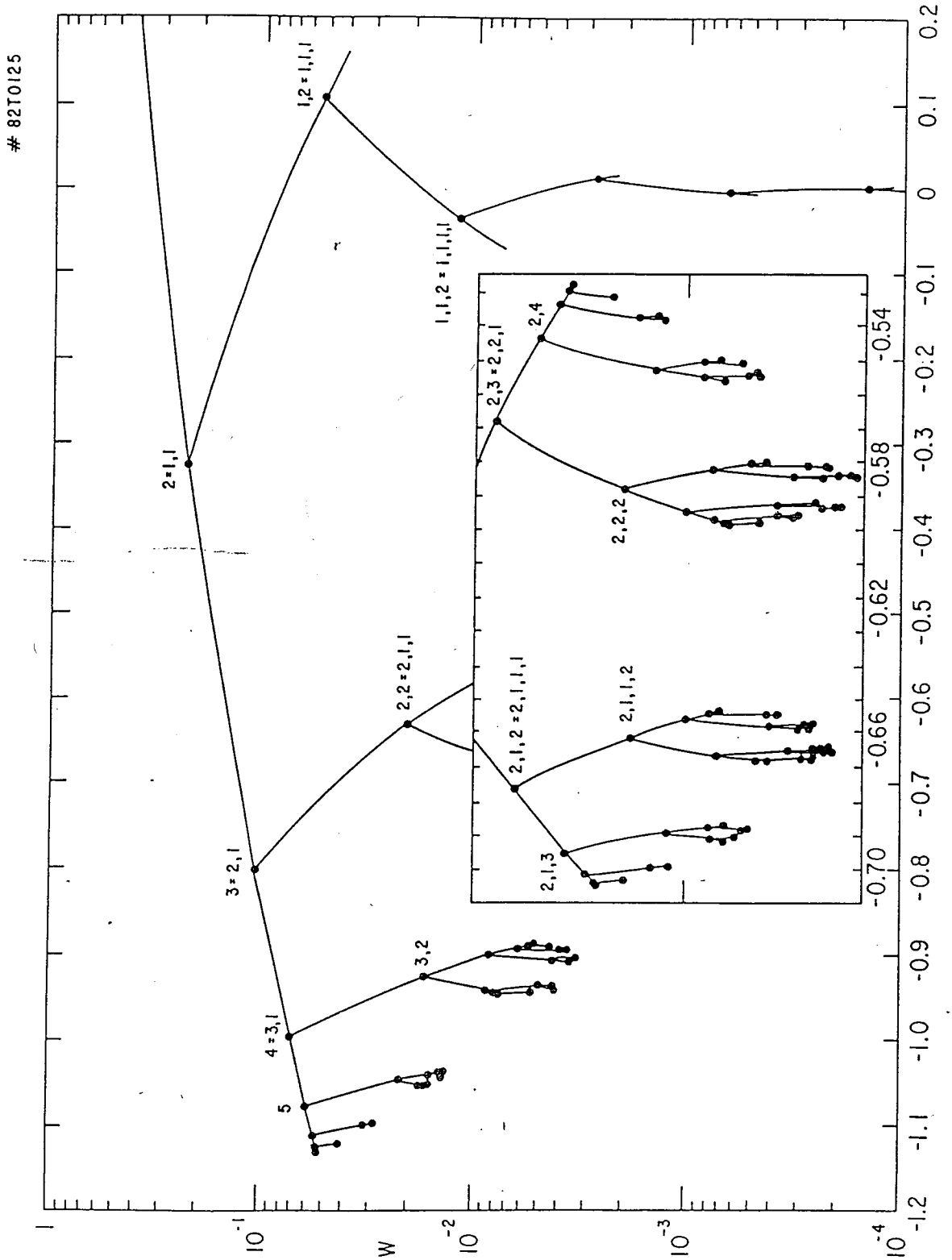
82T0125
ΔW
10^{-1}
10^{-2}
10^{-3}
10^{-4}
1
-1.2
-1.1
-1.0
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
p-p₀
-0.70
-0.66
-0.62
-0.58
-0.54
2=1,1
1,2=1,1,1
1,1,2=1,1,1,1
3=2,1
2,2=2,1,1
4=3,1
3,2
5
2,3=2,2,1
2,4
2,2,2
2,1,2=2,1,1,1
2,1,1,2
2,1,3

FIG. 4.2

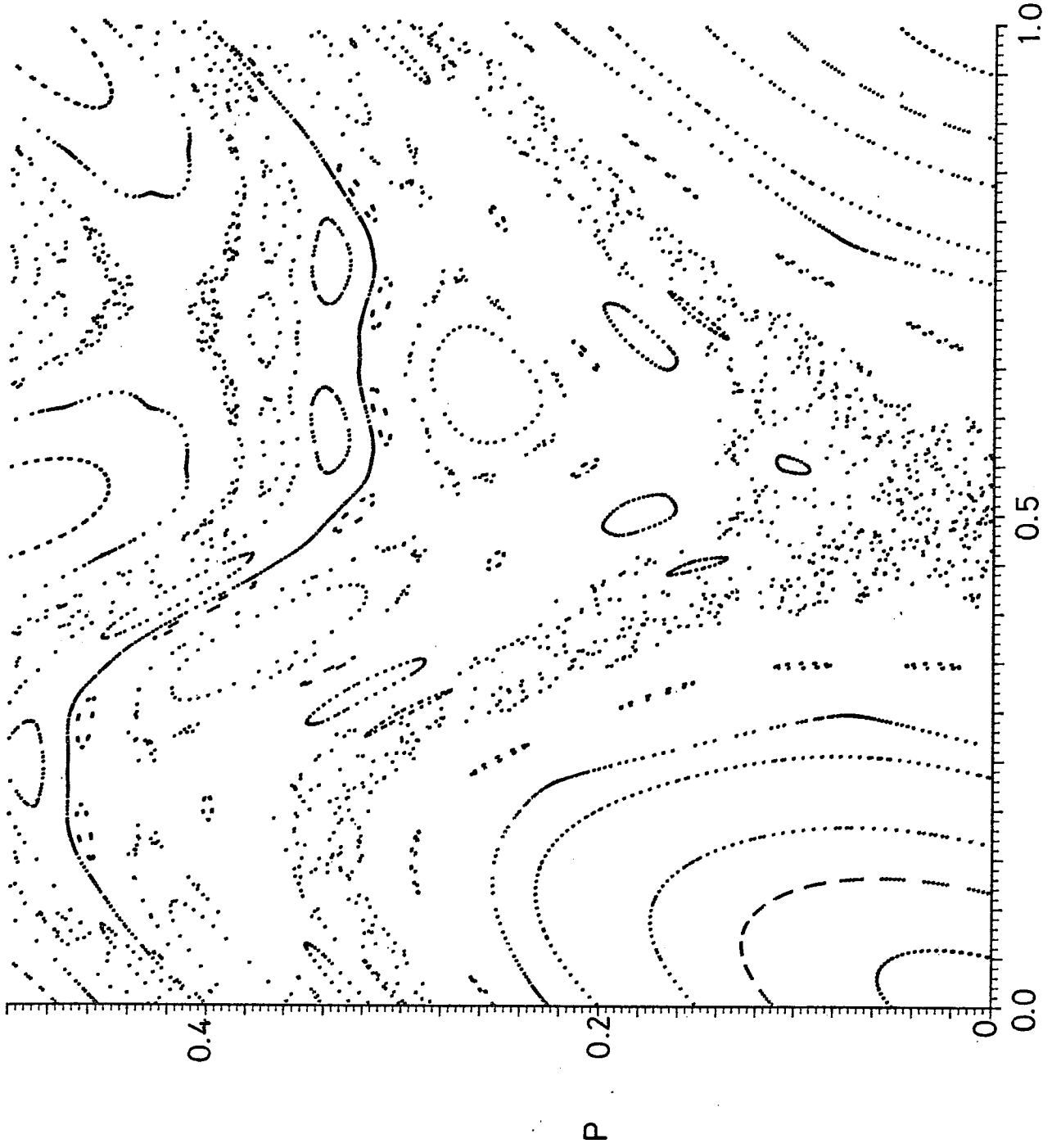

FIG. 5.1

FIG. 6.1

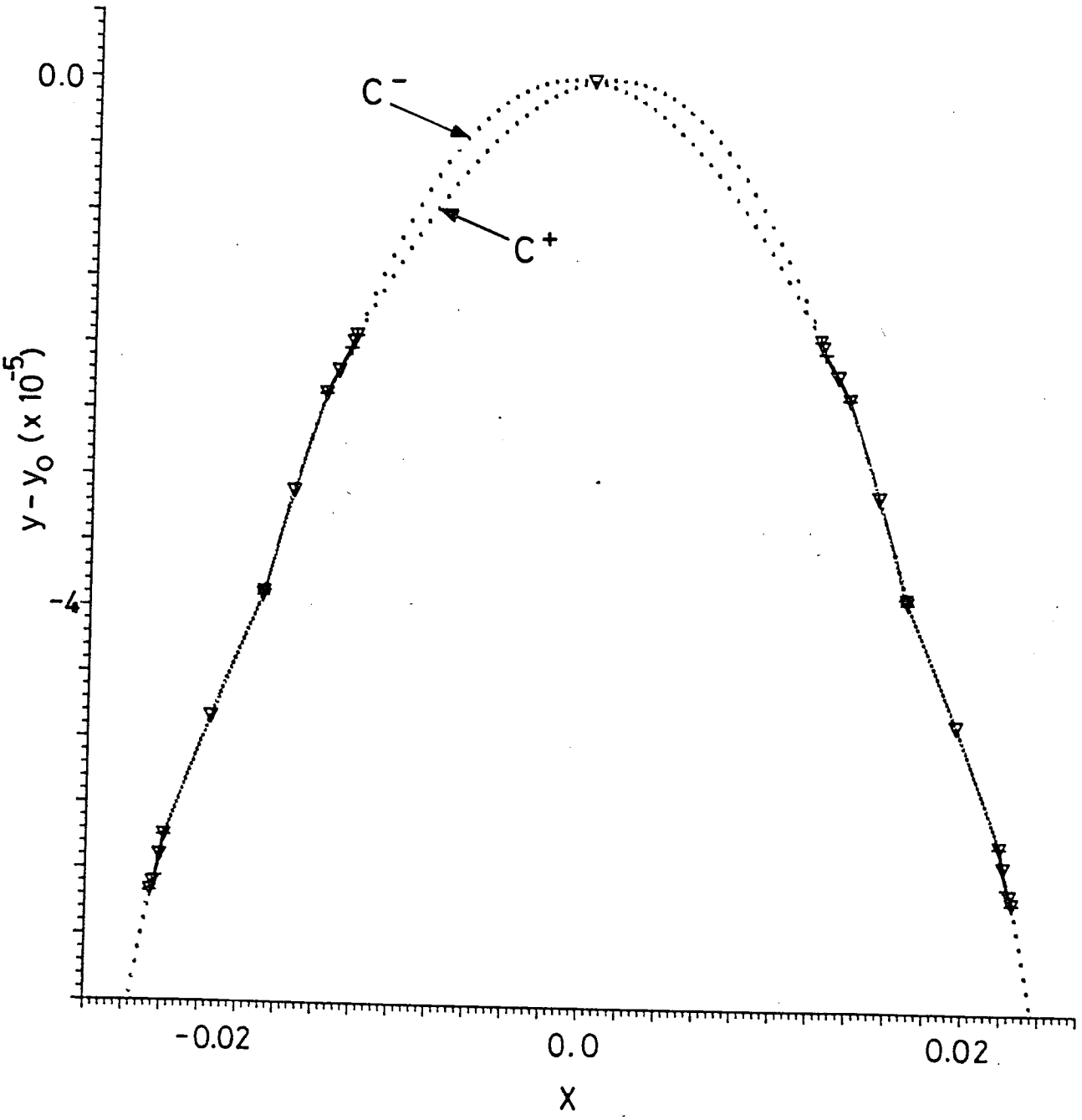

FIG. 6.2

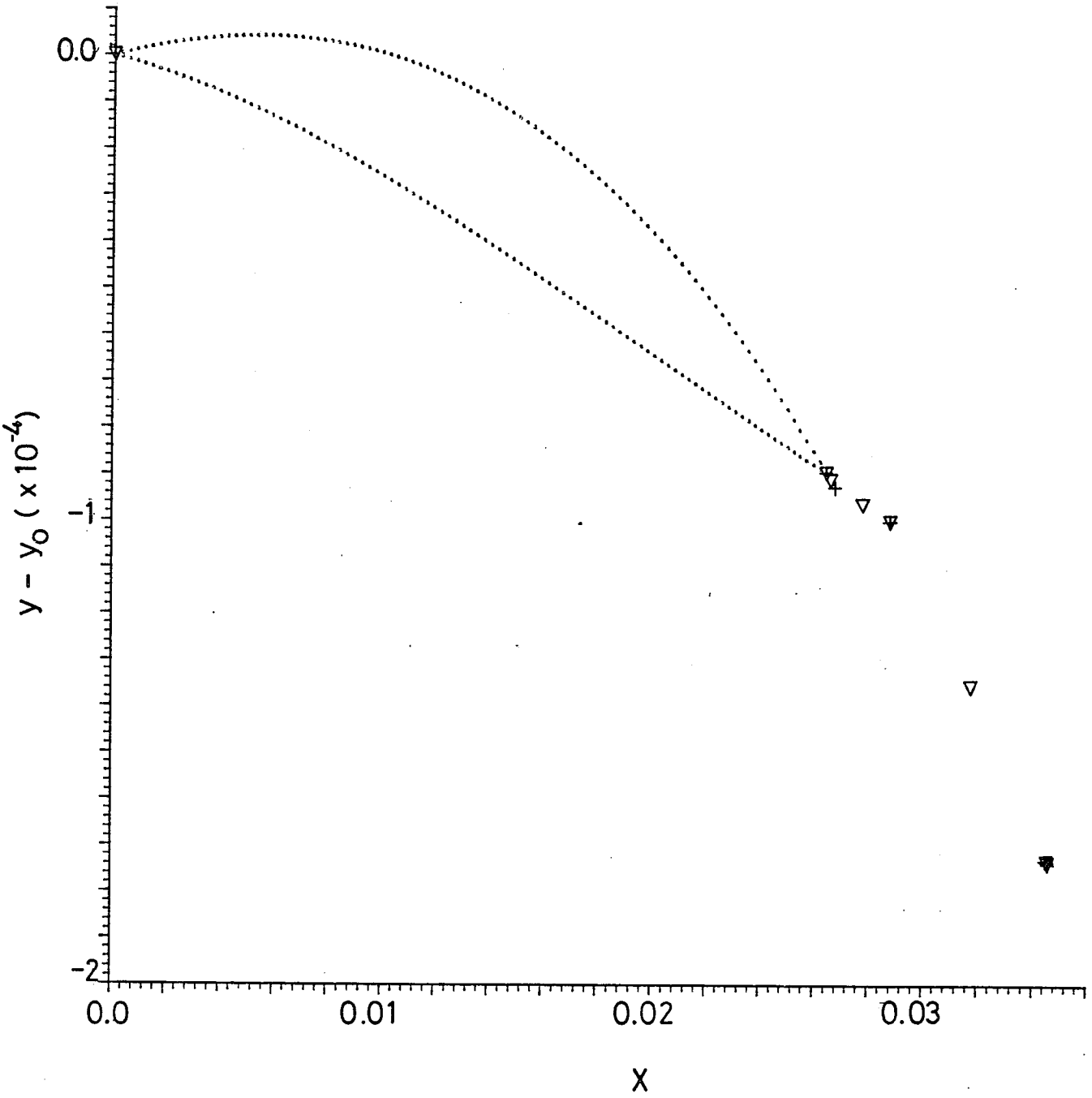

FIG. 6.3a

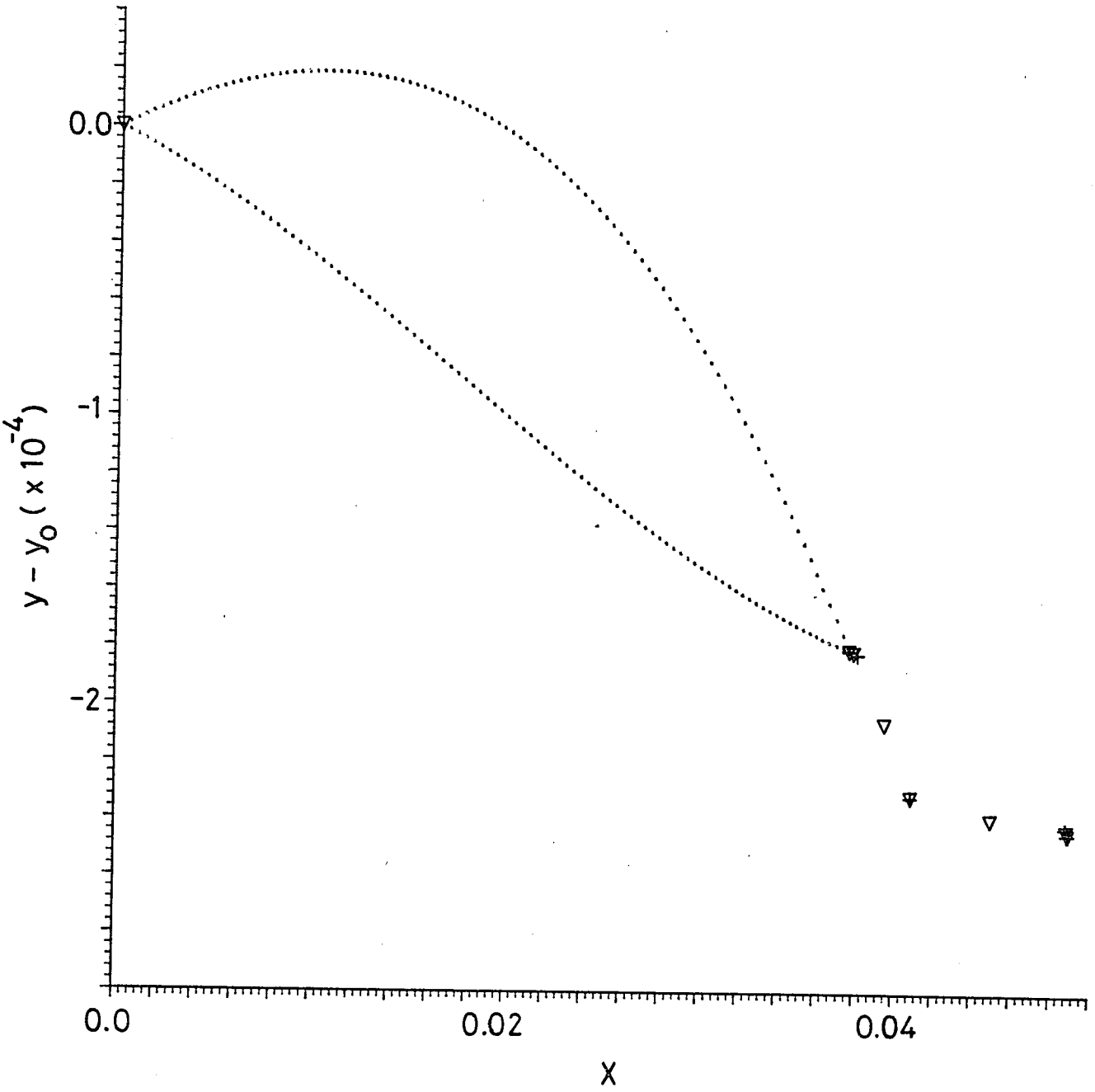

FIG.6.3b

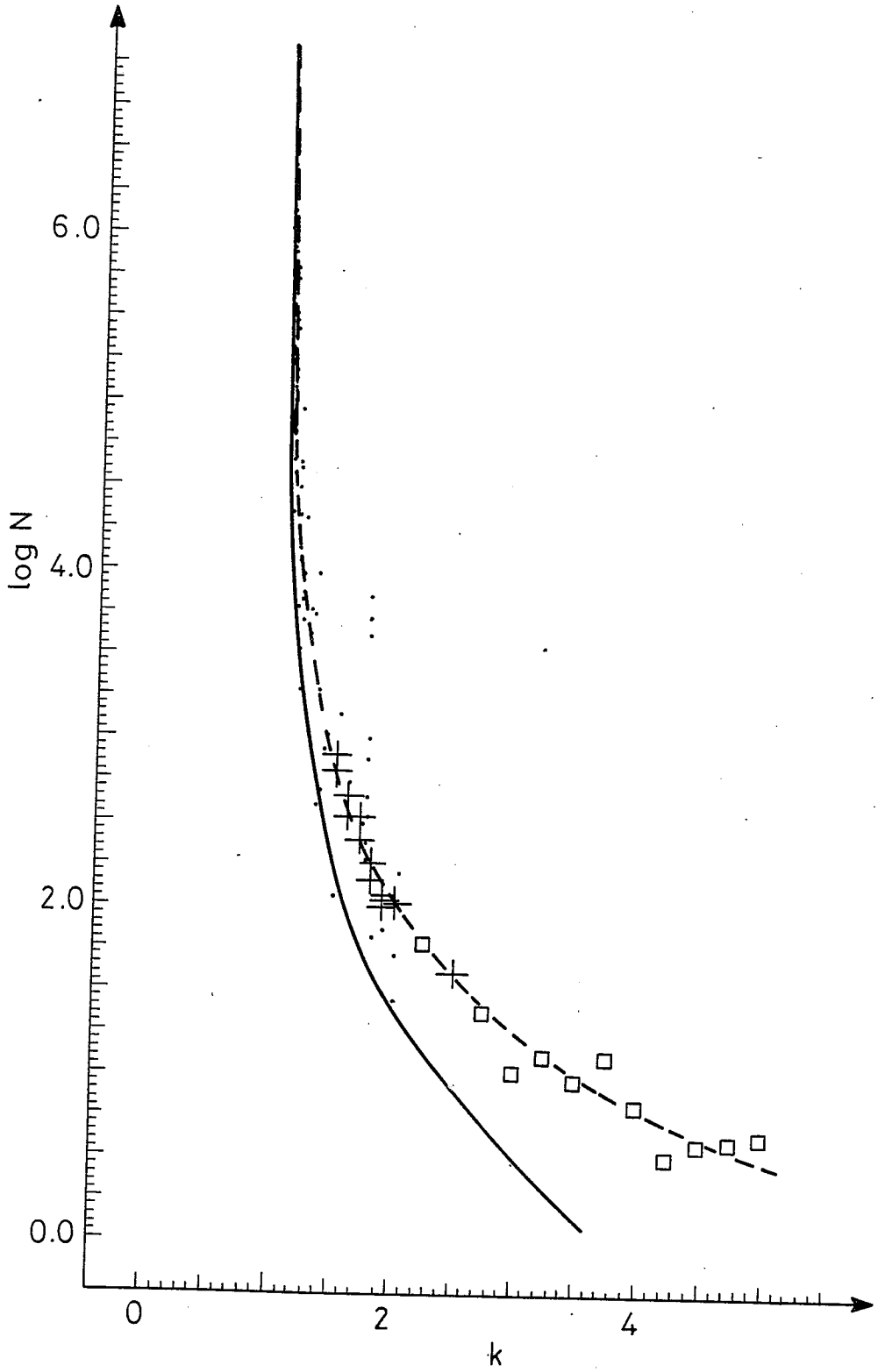

FIG. 11.1