SYMBOLIC DYNAMICS AND MARKOV PARTITIONS

Roy L. Adler
Mathematical Sciences Department
IBM, Thomas J. Watson Research Center
Yorktown Heights, NY 10598

Abstract. The decimal expansion real numbers, familiar to us all, has a dramatic generalization to representation of dynamical system orbits by symbolic sequences. The natural way to associate a symbolic sequence with an orbit is to track its history through a partition. But in order to get a useful symbolism, one needs to construct a partition with special properties. In this work we develop a general theory of representing dynamical systems by symbolic systems by means of so-called Markov partitions. We apply the results to one of the more tractable examples: namely hyperbolic automorphisms of the two dimensional torus. While there are some results in higher dimensions, this area remains a fertile one for research.

1. Introduction.

We address the question: how and to what extent can a dynamical system be represented by a symbolic one? This has roots in a concept of basic arithmetic familiar to us all: namely, the representation of real numbers by infinite binary expansions. As the example of Section 3.2 shows this is related to partitions with special properties which is the theme of this work.

K. Berg [B] in his Ph.D thesis discovered partitions (now called Markov) of the two dimensional torus which give rise to discrete time Markov processes under the action of hyperbolic automorphisms. Shortly after the B. Weiss and the author [AW] proved that such automorphisms are measure theoretically isomorphic if and only if they have the same entropy. Their proof was based on two ideas:

(1) symbolic representations of dynamical systems by means of Markov partitions;
(2) coding between symbolic systems having equal entropy.

Each of these two aspects has undergone extensive development since. The present work is concerned with a systematic treatment of the first idea within the framework of point set topology: that is, we develop the notion of a discrete time topological Markov process without any recourse to measure theory and use it to obtain symbolic representations of dynamical systems. A future paper is planned expanding on the second item above incorporating it in a general isomorphism theory of abstract dynamical systems.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TeX

A disquieting aspect of the result of Adler-Weiss was a certain vagueness where one expects certainty: namely, not quite knowing how to compute the numerical entries of an certain integral matrices proven to exist. Theorem 8.4, the main one of Section 8, is an improvement on the results of Berg and Adler-Weiss, and does not suffer from this difficulty. The proof involves four cases: Case I, the simplest, was done by Anthony Manning many years ago as the author learned from Peter Walters and still may be unpublished.

In Section 2, we briefly introduce the concept of an abstract dynamical system and then go on to give three important concrete examples of such systems: namely multiplication maps, toral automorphisms, and symbolic shifts.

In Section 3, we discuss symbolic representations of dynamical systems and illustrate them for the three concrete systems introduced in the previous section.

In Section 4, we present some general notions needed for our theory of symbolic representation of dynamical systems.

In Section 5, we introduce the notion of topological partition. and show how one gets a symbolic representation from such an object. In order to simplify notation, an improvement in the choice of elements for such partitions was suggested by D Lind: namely replacing "proper sets with disjoint interiors", a *proper* set being one that is the closure of its interior, by "disjoint open sets whose closures cover the space". There is a difference between an open set and the interior of its closure, and exploiting this seemingly slight difference leads not only to notational conveniences but also to pleasant simplifications in subsequent proofs.

In Section 6 we define topological Markov partitions and prove Theorem 6.5, the main theorem of this work. This result concerns getting by means of Markov partitions the best that can be expected as far as symbolic representations of a dynamical systems is concerned. Also in this section we prove a converse to the main theorem, Theorem 6.10, by which one gets Markov partitions from symbolic representations. This leads to the question: does one construct Markov partitions to get symbolic representations or does one produce a symbolic representations to get Markov partitions? The answer to this riddle as far as the current evidence seems to indicate is discussed in Section 9.

In Section 7, we provide results useful for constructing Markov partitions, particularly the final theorem of the section, Theorem 7.12.

In Section 8, we apply Theorem 7.12 to construct certain special Markov partitions for arbitrary hyperbolic automorphisms of the two dimensional torus, which is the content of Theorem 8.4. These partitions have the virtue that a matrix specifying a hyperbolic automorphism is also the one that specifies a directed graph from which the symbolic representation is obtained. The proof we present, though involved, is quite elementary using mainly plane geometry.

The spirit of this work is to rely solely on point set topology. We avoid any measure theory in this discussion. Perhaps a course in point set topology might be spiced up by using items in this work as exercises. In addition, our style of presentation is an attempt to accommodate students as well as experts.

The research behind this work was carried out over many years, in different places, and with help from a number of colleagues, particularly Leopold Flatto and Bruce Kitchens. Work was done at the Watson Research Center, University of Warwick, and MSRI. Most of the research for Sections 5-7 was done in the MSRI 1992 program in Symbolic Dynamics.

2. Abstract and Concrete Dynamical Systems.

At its most simplistic and abstract a dynamical system is a mathematical structure capable of generating orbits which evolve in discrete time. A map of a space into itself will achieve this. Depending on one's purpose additional structure is imposed:

our requires some topology. Consequently, for us an *abstract dynamical system* is a pair (X, ϕ) where X is a compact metric space with metric, say, $d(\cdot\,,\cdot)$ and ϕ is a continuous mapping of a X into itself. We shall refer to X as the *phase* space of the dynamical system. The *orbit* of a point $\mathrm{p} \in X$ is defined to be the sequence $(\phi^n \mathrm{p})_{n=0,1,2,\ldots}$. We shall consider systems where ϕ is onto. Also we shall be mainly, though not exclusively, interested in invertible maps– *i.e.* where ϕ is a homeomorphism–in which case the *orbit* of a point $\mathrm{p} \in X$ is defined to be the bilaterally infinite sequence $(\phi^n \mathrm{p})_{n\in\mathbb{Z}}$. For invertible maps we can speak of past, present, or future points of an orbit depending on whether n is negative, zero or positive; while for non-invertible maps there is only the present and future.

For the above category of abstract systems, we have the following notion of total topological equivalence.

DEFINITION 2.1. Two systems (X, ϕ), (Y, ψ) are said to be *topologically conjugate*, $(X, \phi) \simeq (Y, \psi)$, if there is a homeomorphism θ of X onto Y which *commutes* with ϕ and ψ: *i.e.*, $\phi\theta = \theta\psi$.

We introduce some classical concrete dynamical systems. The first type is most elementary. Though non-invertible, it illustrates admirably some essential ideas which we shall discuss later.

2.1 Multiplication maps.
Let (X, f) be the system whose phase space is the complex numbers of modulus one– *i.e.* elements of the unit circle–acted upon by the mapping $f : z \to z^n$ for some integer $n > 1$.

For our purposes it is more convenient to consider a topologically and algebraically equivalent formulation. Let $X = \mathbb{R}/\mathbb{Z}$ where $\mathbb{R}$ is the real line and $\mathbb{Z}$ the subgroup of integers. Recall that elements in X are cosets modulo $\mathbb{Z}$. The *coset* of $x \in \mathbb{R}$ modulo $\mathbb{Z}$, which we denote by $\{x\}$, is the set $\{x + z | z \in \mathbb{Z}\}$ of *lattice translates* of x. Invoking some standard terminology, the real line $\mathbb{R}$ can be referred to as the *universal cover* of the the circle X. In view of the fact that $\mathbb{Z}$ acts as a group of transformations on the universal cover $\mathbb{R}$, a coset is also called a $\mathbb{Z}-$*orbit*. Two points x, x' in the same coset or $\mathbb{Z}-$orbit are said to be *equivalent* mod $\mathbb{Z}$. The *metric* is given by defining the *distance* between pairs of cosets as the smallest Euclidean distance between pairs of members. Recall that the coset of $x + y$ depends only on the coset of x and that of y: that is, if $x' \in \{x\}$ and $y' \in \{y\}$, then $\{x' + y'\} = \{x + y\}$. Thus *addition* of cosets given by $\{x\} + \{y\} \equiv \{x + y\}$ is well defined and so is the *multiplication-by-n* map $f : \{x\} \mapsto \{x\} + \cdots + \{x\}$ (n-times) $= \{nx\}$. This map is continuous with respect to the metric. It is not invertible: every coset $\{x\}$ has n pre-images which are $\{(x + m)/n\},\ m = 1, \ldots, n$.

The closed unit interval $[0, 1]$ is a set referred to as a fundamental region for the action of $\mathbb{Z}$ on $\mathbb{R}$.

DEFINITION 2.1.1. A *fundamental region* is defined as a closed set such that

(1) it is the closure of its interior;
(2) every orbit under the action has at least one member in it (this is equivalent to the statement that the translates of the unit interval by elements of $\mathbb{Z}$ tile $\mathbb{R}$);

(3) no point in the interior is in the same $\mathbb{Z}$-orbit as another one in the closed region (this restriction does not apply to two boundary points– *e.g.* the points 0 and 1 are in the same one).

Fundamental regions are not unique: for example, the interval $[1, 2]$ is also a fundamental region for the action of $\mathbb{Z}$ on $\mathbb{R}$, though not a particularly useful one.

One can give another equivalent reformulation of the phase space of these systems in terms of a fundamental region with boundary points identified. Let X be the closed unit interval with 0 identified with 1. We define a metric on X by

$$d(x, y) = \min(|x - y|, |x - y - 1|, |x - y + 1|).$$

From now on let us take the notation $\{x\}$ to mean fractional part of a real number x. On X the map f takes the form

$$f(x) = \{nx\}.$$

Since all numbers in a coset have the same fractional part and that number is the unique member of the intersection of the coset and X, this new interpretation of $\{x\}$ is consistent with the old.

A more serious set of examples are continuous automorphisms of certain compact Abelian groups–namely the n-dimensional tori. For simplicity we restrict the discussion to the case of dimension two, generalization to higher dimensions being quite analogous.

2.3 Toral Automorphisms.
Consider the two dimensional torus $\mathbb{R}^2/\mathbb{Z}^2$ and a continuous group automorphism ϕ. Here the universal cover of the 2-torus is $\mathbb{R}^2$. The description of the action of the integers on the real line generalizes in a straight-forward manner to the action of the subgroup $\mathbb{Z}^2$ of points with integer coordinates on the universal cover $\mathbb{R}^2$. The definitions of cosets, lattice translates, orbits, addition, and the metric are quite similar.

A continuous automorphism ϕ is specified by a a 2×2 matrix A : with integer entries and determinant ± 1. Let

$$\mathrm{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

The matrix A determines an invertible linear transformation on $\mathbb{R}^2$. We represent the points in the plane by row vectors and the action of the linear transformation by right[1] matrix multiplication. The map ϕ is then defined as follows: the image of the coset containing (x, y) is the one containing $(ax + cy, bx + dy)$. This map is well-defined– *i.e.*, the image does not depend on the choice of coset representative (x, y)– because A is invertible and maps $\mathbb{Z}^2$ onto itself.

Some things to note. The map ϕ is a homeomorphism. The coset $\{0\} = \mathbb{Z}^2$ is a fixed point of ϕ. There may of course be other fixed points. A pair (x, y) is in a

[1]Right multiplication on row vectors turns out to be a little more convenient than left multiplication on column vectors.

coset which is a fixed point if and only if it satisfies the two linear equations given by

$$(x, y)\mathrm{A} = (x, y) + (m, n)$$

for some pair of integers (m, n). The only solutions are rational. In addition, a coset is periodic under ϕ if and only if it is fixed under some iterate A^n. Furthermore, if a coset contains a point with rational coordinates, then it is periodic which follows from the fact that the product of the denominators in a rational pair (x, y) bounds the denominators in the sequence $(x, y), (x, y)\mathrm{A}, (x, y)\mathrm{A}^2, \ldots$ which implies that the orbit of (x, y) is finite.

The plane $\mathbb{R}^2$ is the universal cover of the 2-torus, and any closed unit square with lattice points as corners is a fundamental region. We shall call the one with its lower left corner at the origin the *principal fundamental region*. Like the one dimensional case, we can formulate the system in terms of it. Let X be the closed unit square with each point on one side identified with its opposite on the other. The coset of (x, y) intersects X in a unique point: namely, $(\{x\}, \{y\})$. On X the map ϕ takes the form

$$\phi(x, y) = (\{ax + cy\}, \{bx + dy\}).$$

Unlike the case of one dimension, other fundamental regions, as we shall see, play a crucial role in studying the action of automorphisms.

Finally we come to the most basic of concrete systems. Ultimately we shall show to what extent they model others, in particular multiplication maps and hyperbolic toral automorphisms.

2.4 Symbolic Shifts[2].

Let $\mathcal{A}$, called an *alphabet*, denote an ordered set of N symbols, often taken to be $\{0, 1, \ldots, N-1\}$. The domain of this system is the space

$$\Sigma_N = \mathcal{A}^{\mathbb{Z}} = \{s = (s_n)_{n\in\mathbb{Z}} | s_n \in \mathcal{A}\} \tag{2.4.1}$$

of all bi-infinite sequences of elements from a set of N symbols. One can think of an element of this space as a bi-infinite walk on the complete directed graph of N vertices which are distinctly labelled. Sometimes it is more convenient to label edges, in which case the picture is a single node with N oriented distinctly labelled loops over which to walk. In Figure 2.I we have depicted the full 2-shift by both types of graph labelling.

The *shift transformation* σ is defined by shifting each bi-infinite sequence one step to the left. This is expressed by

$$(\sigma s)_n = s_{n+1}.$$

We define the distance $d(s, t)$ between two distinct sequences s and t as $1/(|n| + 1)$ where n is the coordinate of smallest absolute value where they differ. Thus if $d(s, t) < 1/n$ for $n > 0$ then $s_k = t_k$ for $-n < k < n$. This metric makes the space

[2]For comprehensive treatment of this area we refer the reader to the book by Lind and Marcus [LM].

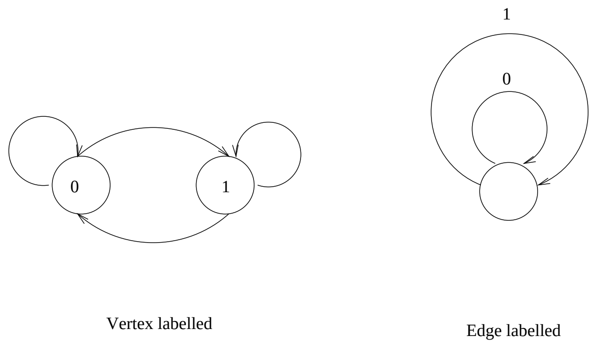


FIGURE 2.I. Full 2-shift

Σ_N one of the important compact ones–namely, the Cantor discontinuum– and the shift a homeomorphism. The symbolic system (Σ_N, σ) is called the *full N-shift*.

Restricting the shift transformation of a full shift Σ_N to a closed shift-invariant subspace Σ, we get very general kind of dynamical system (Σ, σ) called a *subshift.* Given a symbolic sequence $s = (s_n)_{n\in\mathbb{Z}}$ and integers $m < n$, we shall use the notation $s_{[m,n]}$ to stand for the $m-n+1$-tuple $(s_m, s_{m+1}, \dots, s_n)$. Given a symbolic phase space Σ, we call a k-tuple an *allowable k-block* if it equals $s_{[m,m+k-1]}$ for some $s \in \Sigma$.

Returning to the realm of the more specific from our momentary excursion into the less knowable, we define *shift of finite type*, also called *topological Markov shift* , as the subshift of a full shift restricted to the set Σ_G of bi-infinite paths in a finite directed graph G derived from a complete one by possibly removing some edges.

Usually we denote the space Σ_G by Σ_{A} where A is a matix of non-negative integers a_{ij} denoting the number of edges leading from the i-th node to the j-th. The term "Markov" is derived from the resemblance to Markov chains for which the a_{ij} are probabilities instead of integers. One thing to note is that the ij entry of A^n is the number of paths of length n beginning at $i-$th node and ending at the $j-$th. Often however A is an $N \times N$ matrix of zeroes and ones specifying a directed graph of N nodes (edges) according to the following: the i-th node (edge) is connected to the j-th, $i \longrightarrow j$, if and only if $a_{ij} = 1$. Whether dealing with nodes or edges, we call A a *transition* matrix and restate for zero-one matrices the above definition by

$$\Sigma_G = \Sigma_{\mathrm{A}} \equiv \{s = (\dots, s_n, \dots) | a_{s_n, s_{n+1}} = 1, s_n \in \mathcal{A}, n \in \mathbb{Z}\}. \tag{2.4.2}$$

REMARK.. Let (Σ_G, σ) be a topological shift given by a node-labelled directed graph G. Nodes from which there is no return are called *transient*, the rest *recurrent.* A node is transient if and only if either it has no predecessor nodes or all its predecessors are transient. This statement is not as circular as it seems: for the set of predecessors of any set of transient nodes, if non-empty, is a strictly smaller set of transient nodes. The only symbols which appear in bi-infinite sequences of Σ_G are labels of recurrent nodes.

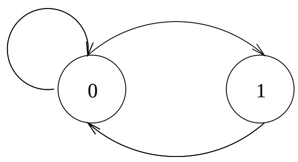


FIGURE 2.II. Fibonacci shift

Figure 2.II describes the *Fibonacci* or *golden ratio* shift, so-called because the number of admissible n-blocks (paths of length n) are the Fibonacci numbers– namely, there are two 1-blocks, three 2-blocks, five 3-blocks, ...

Here the space Σ_{A} is given by the matrix

$$\mathrm{A} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}.$$

If we label the first node by 0 and and the second by 1, then only sequences of 0's and 1's with 1's separated by 0's are admissible. While other shifts of finite type can be specified by graphs with either nodes or edges labelled, there is no edge labelled graph for the Fibonacci shift.

Given a node labelled graph G, we define the *edge graph* $G^{(2)}$ by labelling the edges. For labels we can use the allowable 2-blocks. In general we define the *higher edge graphs* $G^{(n)}$ as follows. The alphabet consists of all allowable blocks $[a_1, \ldots, a_n]$ gotten from paths of length n on G. The transitions are defined by

$$[a_1, \ldots, a_n] \longrightarrow [b_1, \ldots, b_n]$$

if and only if $b_1 = a_2, \ldots, b_{n-1} = a_n$.

There is a one-side version of the full N-shift: namely,

$$\Sigma_N^+ = \{s = (s_0, s_1, \ldots) | s_n \in \mathcal{A}, n = 0, 1, 2, \ldots\}. \tag{2.4.3}$$

On this space the shift transformation σ is similarly defined: namely, $(\sigma s)_n = s_{n+1}$ but only for non-negative n. It acts by shifting sequences one step to the left and dropping the first symbol. The metric on this phase space is defined the same as before but absolute value signs are not needed. We also have one-side versions of shifts of finite type. In one-side symbolic systems, the shift transformation, like a multiplication map, is continuous but not invertible.

Exercises.

3.1 Prove that the canonical map $\psi : G^{(n)} \longrightarrow G$ defined by $\psi[s_1, \ldots, s_n] = s_1$ gives topological conjugacy of $(\Sigma_{G^{(n)}}, \sigma)$ and (Σ_G, σ).

3 Symbolic Representations.

Shifts of finite type contain a great deal of complexity, yet are the best understood dynamical systems. Such symbolic dynamical systems can be used to analyze general discrete time ones. For example, a good symbolic representation will show how to identify periodic orbits, almost periodic ones, dense ones, etc..

Representing a general dynamical system by a symbolic one involves a fundamental complication. We have two desires: we would like a continuous one-to-one correspondence between orbits $\phi^n x$ of the first and orbits $\sigma^n s$ of the second; and we want the shift system to be one of finite type. Unfortunately these two desires are in conflict: constraints placed by topology must be observed. On one hand a continuous one-to-one correspondence makes X homeomorphic to a shift system. On the other hand a shift system is totally disconnected while X is often a smooth manifold. Thus for the most part we must abandon the quest of finding a topological conjugacy between a given dynamical system and a shift of finite type. However, we shall see that by sacrificing one-to-one correspondence we can still salvage a satisfactory symbolization of orbits. We are reminded of arithmetic in which we represent real numbers symbolically by decimal expansions, unique for the most part, but must allow two expansions for certain rationals. To do otherwise would just make the instructions for arithmetical operations unnecessarily complicated. The most natural way to associate a symbolic sequence with a point in a dynamical system is to track its history as illustrated in Figure 3.I through a family of sets indexed by an alphabet of symbols.

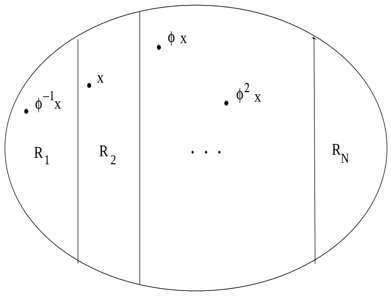


FIGURE 3.I. Partitioning a dynamical system

This is easy, but what is more difficult is to get a family for which each history represents just one point. It is no achievement to specify a family for which each history might represent more than one point. However, we must live with the inevitability that each point might have more than one associated histories. Having found a family of sets, the orbits through which determine a unique point, we want still more: namely, we would like the totality of sequences which arise to comprise a subshift of finite type. In order to do this, we must find a family with special properties. We shall look at some examples for guidance as to what these properties ought to be, and families of sets possessing them will be called *Markov* partitions.

The first example is the trivial case of a dynamical system which is identical with its symbolic representation: namely, a topological Markov shift.

3.1 Cylinder Set Partition for Symbolic Sequences.
Let (Σ_A, σ) be a topological Markov shift, vertex labelled by an alphabet $\mathcal{A}$. We form the partition $\mathcal{C} = \{C_a : a \in \mathcal{A}\}$ of elementary cylinder sets determined by fixing the 0-th coordinate: *i.e.*, $C_a = \{s \in \Sigma_A : s_0 = a\}$. Tracking the history of an

orbit of a element $s \in \Sigma_A$ through this partition means getting a sequence $(s_n)_{n\in\mathbb{Z}}$ such that $\sigma^n s \in C_{s_n}$. But this sequence is s itself. Let us point out the salient features of this partition.

First[3],

$$\bigcap_{n=0}^{\infty}\bigcap_{-n}^{n}\phi^{-k}C_{s_k} = \{s\}$$

Second, if $s \in C_a \cap \sigma^{-1}C_b \neq \emptyset$ then $s \in C_a$ and $\sigma s \in C_b$: *i.e.*, $s_0 = a, s_1 = b$. In terms of the graph, this means there is an edge from a to b. An absolutely obvious property of directed graphs is the following. If there is an edge from a to b, and an edge from b to c, then there is a path from a to c via b. This property can be reformulated as follows. If $C_a \cap \sigma^{-1}C_b \neq \emptyset$ and $C_b \cap \sigma^{-1}C_c \neq \emptyset$ then $C_a \cap \sigma^{-1}C_b \cap \sigma^{-2}C_c \neq \emptyset$. This property has a length n version for arbitrary n: namely, n abutting edges form a path of length $n+1$ and this can be reformulated to read that n pair-wise non-empty intersections lead to an $(n+1)-$fold non-empty intersection. We shall call such a countable set of conditions for $n = 2, 3, ...$ the *Markov* property: it turns out to be a key requirement in getting the desired symbolic representation from a partition.

Finally, there is another important feature of the partition $\mathcal{C} = \{C_a : a \in \mathcal{A}\}$: namely, the sets of this partition have a product structure respected by the shift which is described as follows. Let $s \in C_a$ – in other words, $s_0 = a$ – and define two sets

$$v_a(s) \equiv \bigcap_{0}^{\infty}\sigma^{-k}C_{s_k}$$

which we shall call the *vertical through* s and

$$h_a(s) \equiv \bigcap_{-\infty}^{0}\sigma^{-k}C_{s_k}$$

which we shall call the *horizontal.* A sequence $s \in C_a$ is the sole member of the intersection of its vertical and horizontal– *i.e.* $\{s\} = v_a(s) \cap h_a(s)$. Furthermore, for $s, t \in C_a$ there is a unique sequence in the intersection of the horizontal through s and the vertical through t : namely, $\{(\dots s_{-2}, s_{-1}, s_0 = t_0, t_1, t_2, \dots)\} = v_a(s) \cap h_a(t)$. We define a map of $C_a \times C_a$ onto C_a by $(s,t) \mapsto h_a(s) \cap v_a(t)$, or rather the sole element of this intersection. It is easily verified that this map is continuous and its restriction to $h_a(s) \times v_a(t)$ for any $s, t \in C_a$ is a homeomorphism of $h_a(s) \times v_a(t)$ onto C_a. Finally σ respects this product structure in the sense that if $s \in C_a \cap \sigma^{-1}C_b$ then:

$$\sigma v_a(s) \subset v_b(\sigma s),$$

$$\sigma h_a(s) \supset h_b(\sigma s).$$

This last property is closely connected with the Markov one.

The next example is based on the binary expansions of real numbers, and illustrates what one should expect of a good symbolic a representation of a dynamical system.

[3]From now on we shall commit a convenient semantic error of confusing a set consisting of a single point with the point itself and so dispense with the surrounding braces

3.2 Symbolic Representation for Multiplication by Two.
Let (X, f) be the multiplication system where and $f : x \to \{2x\}$. Recall that the domain Σ_2^+ of the one-sided full 2-shift dynamical system $(\Sigma_{[2]}^+, \sigma^+)$ is the set of one-sided infinite walks on the edge-labelled graph in Figure 2.I. We can equate a sequence $s = (s_n)_{n\in\mathbb{Z}}$ with the binary expansion $.s_1 s_2 s_3 \ldots$. Consider the map π from $\Sigma_{[2]}^+$ to X defined by $\pi(s_1, s_2, \ldots) = \{s_1/2 + s_2/4 + \ldots\}$. It is readily verified that

(i) $f\pi = \pi\sigma^+$,
(ii) π is continuous,
(iii) π is onto,
(iv) there is a bound on the number of pre-images (in this case two),

and

(v) there is a unique pre-image of "most" numbers (here those with binary expansions not ending in an infinite run of all zeros or all ones).

The map π is not a homeomorphism, but we do have a satisfactory representation of the dynamical system by a one-sided 2-shift in the sense that: orbits are preserved; every point has at least one symbolic representative; there is a finite upper limit to the number of representatives of any point; and every symbolic sequence represents some point. This is a example of what is known as a *factor map* which we shall formalize in §4.

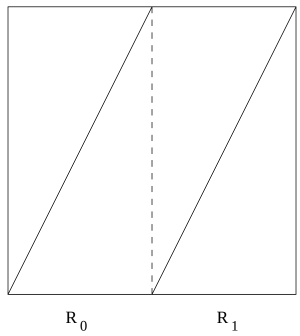


FIGURE 3.II. Graph of multiplication by 2 (mod 1)

As we have led to reader to expect, there is an alternate definition of π in terms of a partition. Consider $\mathcal{R} = \{R_0 = (0, 1/2), R_1 = (1/2, 1)\}$. The elements of this family are disjoint open intervals whose closure cover the unit interval. The map π which associates sequences with points has an alternate expression in terms of this family: namely,

$$\pi(s_1, s_2, \ldots) = \bigcap_{n=0}^{\infty} \overline{R_{s_1} \cap f^{-1}(R_{s_2}) \cap \cdots \cap f^{-n}(R_{s_{n+1}})}.$$

Remark. The reader might wonder about defining π by the simpler expression

$$\pi(s_1, s_2, \dots) = \bigcap_{n=0}^{\infty} \overline{f^{-n}(R_{s_{n+1}})}.$$

There are cases where this would suffice, but a difficulty can arise and does here. In X the point 0 which is identified with 1 is a fixed point of f which implies that $0 \in f^{-n}\overline{R_i}$ for $i = 0, 1$ and $n \geq 0$. Thus, except for the all 0 or all 1 sequence, $\pi(s_1, s_2, \dots)$ is a set which does not consist of a singleton: it contains two real numbers, one of which is the fixed point 0; and this renders π ill-defined. The most we can say in general is that

$$\bigcap_{n=0}^{\infty} \overline{R_{s_1} \cap f^{-1}(R_{s_2}) \cap \cdots \cap f^{-n}(R_{s_{n+1}})} \subsetneq \bigcap_{n=0}^{\infty} \overline{f^{-n}(R_{s_{n+1}})}$$

However, for the so-called expansive dynamical systems, when the size of partition elements is uniformly small enough, equality holds in which case π would be well-defined (see Proposition 5.8).

Next we consider hyperbolic automorphisms of the 2-torus. This was the first smooth class of invertible dynamical systems found to have Markov partitions. This discovery was made by K Berg [B] in 1966 in his doctoral research. A short time later R. Adler and B Weiss [AW] constructed some special Markov partitions in order to prove that two such systems are conjugate in the measure theoretic sense if they have the same entropy. For these systems topological conjugacy implies measure conjugacy, but not conversely. We shall give a formal development the general two-dimensional case in a later chapter. Before making that plunge, we shall wet our toes with an informal discussion of one specific illustrative case. A rigorous proof of what we are about to describe will be achieved by Theorem 7.13.

3.3 Partition for a Toral Automorphism.
Take the matrix

$$\mathrm{A} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

which we have met before in quite a different context. Let X be the two-torus and ϕ be given by A : that is,

$$\phi(x, y) = (\{x + y\}, x).$$

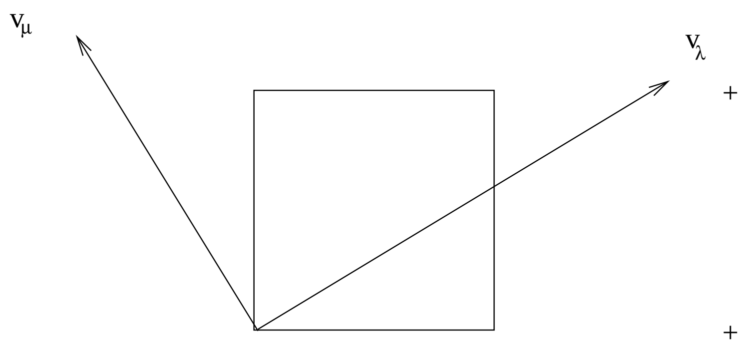


FIGURE 3.III. The torus and eigen-directions of A

The matrix A has two eigenvalues: $\lambda = (1+\sqrt{5}\,)/2$ and $\mu = (1-\sqrt{5}\,)/2$. Observe that $\lambda > 1$ and $-1 < \mu < 0$. Associated with these eigenvalues are the eigenvectors v_λ pointing into the first quadrant and v_μ into the second. In Figure 3.III we have drawn two lines through the origin in the eigenvector directions. The action of A on a vector is to contract its $\mathbf{v}_\mu$-component by $|\mu|$ and expand its v_λ-component by λ. Note μ is negative, which causes a direction reversal besides a contraction in the v_μ-component. We refer to the direction of v_λ as the *expanding* direction and that of v_μ as the *contracting* direction.

In Figure 3.IV we draw another region with sides parallel to the expanding and contracting directions. That it is a fundamental region is verified by noting that each of the three triangles sticking out of the unit square is a translation by an element of $\mathbb{Z}^2$ of one of the three missing triangles inside.

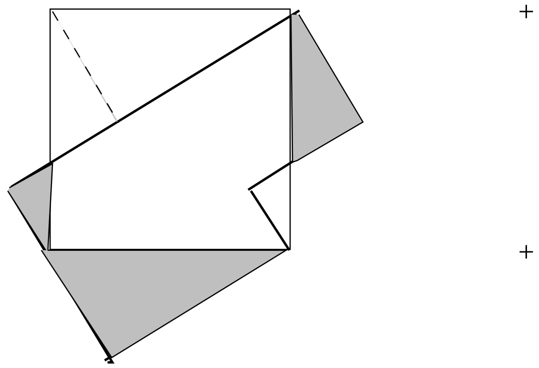


FIGURE 3.IV. Another fundamental region of 2-torus

We call this fundamental region the *principal* one, and draw within it the collection of open rectangles $\mathcal{R} = \{R_i \ : i = 1,2,3\}$ as depicted in Figure 3.V. This family is an example of a type of partition we shall later describe as *Markov.* We label significant points using the same letters for those which are equivalent.

The image of this partition under the linear transformation determined by A is depicted in Figure 3.VI. In drawing it the following calculations come into play: $(1,0)\mathrm{A} = (1,1),\ (1,1)\mathrm{A} = (2,1),\ (2,1)\mathrm{A} = (3,2)$.

Along with the image we have included an outline of the original partition. Notice that ϕR_3 is actually the same as R_2. Also notice how the other ϕR_i overlap R_1 and R_3. The manner in which the image partition intersects the original partition can be summed up as follows: $\phi R_i \cap R_j \neq \emptyset$ according to whether j follows i in the edge graph in Figure 3.VII.

The boundary of the sets in $\mathcal{R}$, $\partial R_i = \overline{R_i} - R_i$, consists of various line segments in the v_λ and v_μ directions. The union of those of the R_i's in the v_λ-direction is called the *expanding boundary of the partition* and those in the v_μ-direction, the *contracting boundary.* By lattice translations of the various bounding segments, we

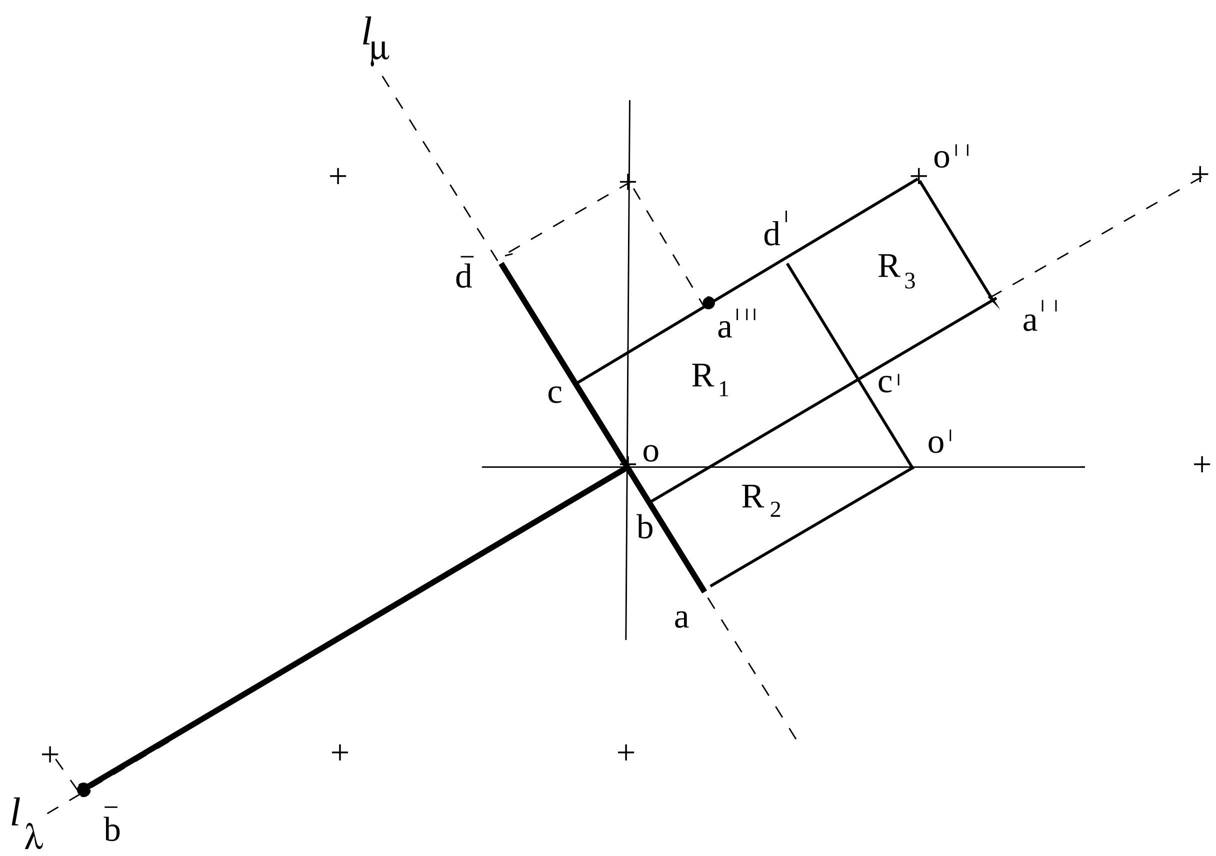


FIGURE 3.V. Partition of 2-torus

can reassemble their union into two intersecting line segments through the origin, $o\bar{b}$ and $a\bar{d}$, as shown in Figure 3.VI.

The behavior of the boundary under the action A leads to a topological Markov shift representation. The essential properties are that $a\bar{d}$ contains its image under A; whereas $o\bar{b}$ is contained in its image, or equivalently $o\bar{b}$ contains its inverse image. Because A preserves eigen-directions and keeps the origin fixed, it is easy to see that $o\bar{b}$ gets stretch over itself; but because there is a reflection involved it is not enough to know that the length of $a\bar{d}$ is contracted by A. We must show that

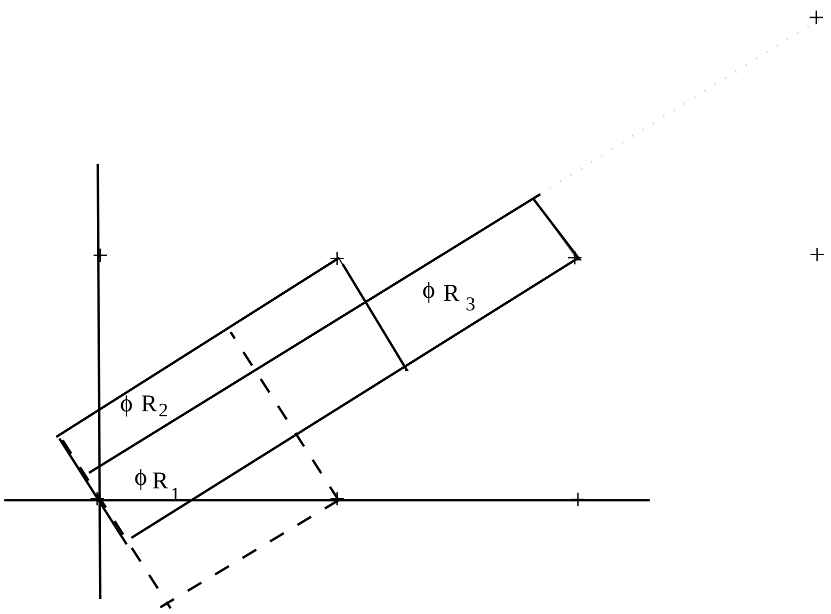


FIGURE 3.VI. Its image

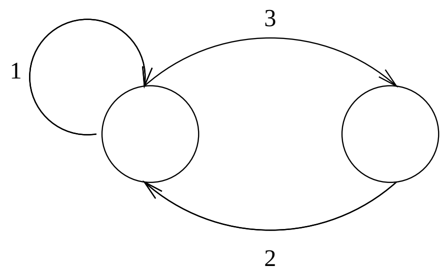
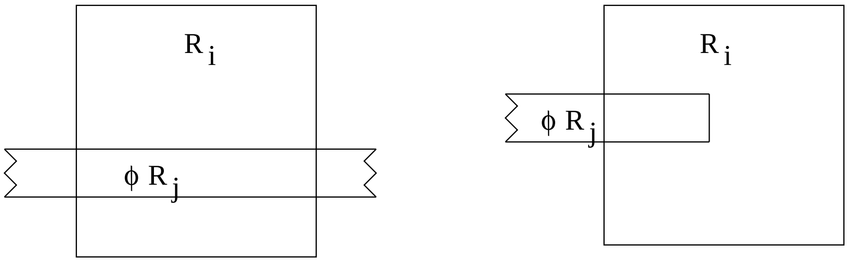


FIGURE 3.VII. Edge graph for ϕ acting on $\mathcal{R}$

the points a and $\bar{\mathrm{d}}$ on the line segment a$\bar{\mathrm{d}}$ have their images within that segment. These points are the projections to l_μ in the $\mathbf{v}_\lambda$-eigen-direction from $(0,1)$ and $(1,0)$ respectively: so their images are the projections from the images of these lattice points which are $(1,0)$ and $(1,1)$ respectively. Thus the image of $\bar{\mathrm{d}}$ is a, and the image of a is c. These facts about the expanding and contracting boundaries imply that refinements of the original partition under positive iterates of ϕ do not have any new boundary segments in the v_μ-direction that aren't already contained in a$\bar{\mathrm{d}}$, while under negative iterates of ϕ there are no new ones in the v_λ-direction not already in $o\bar{\mathrm{b}}$. From this we obtain that for $n \in \mathbb{N}$ a set $\phi^n R_i \cap R_j$, if non-empty, is a union of rectangles stretching in the expanding direction all the way across R_j. Similarly, a non-empty $R_i \cap \phi^{-n} R_j$ is a union of rectangles stretching in the contracting direction all the way across R_i. When $n = 1$, it can be seen that these each of these unions consists of a single rectangle. This implies that if

$\bigcap_{k=-n}^{0} \phi^{-k} R_{s_k} \neq \emptyset$, then this intersection is a single rectangle stretching all the way across R_{s_0} in the expanding direction. Similarly, if $\bigcap_{k=0}^{n} \phi^{-k} R_{s_k} \neq \emptyset$, then this set is a single rectangle stretching all the way across R_{s_0} in the contracting direction.



Possible Impossible

FIGURE 3.VIII. Intersections

Combining these two results we have that a non-empty closed set of the form $\overline{\bigcap_{k=-n}^{n} \phi^{-k} R_{s_k}}$ is a closed rectangle. The diameter of these sets is uniformly bounded by constant $\times\ |\mu|^n$. Thus as $n \to \infty$, a sequence of such sets decreases to a point in X. Consequently, such a point can be represented by a sequence $s = (s_n)_{n \in \mathbb{Z}}$. If fact, all points of the torus can be so represented.

If $\bigcap_{k=-n}^{n} \phi^{-k} R_{s_k} \neq \emptyset$, then it is clear that $R_{s_i} \cap \phi^{-1} R_{s_{i+1}} \neq \emptyset$ for $-n \leq i \leq n-1$. The converse which is the Markov property is really the main one we are extracting from the geometry of this example. As we have seen $R_{s_i} \cap \phi^{-1} R_{s_{i+1}} \neq \emptyset$ if and only if edge s_{i+1} follows edge s_i according to the the graph of Figure 3.VII. This means that the sequences $s = (s_n)_{n \in \mathbb{Z}}$ are elements is a topological Markov shift.

Once again sets R_i have an obvious product structure. For p $\in R_i$ we call the segment h_i(p) specified by intersection of R_i and the line through p in the expanding direction the *horizontal* through p. Similarly, we refer to v_i(p) given by the intersection of R_i and the line through p in the contraction direction as the *vertical*. Each rectangle is homeomorphic to the Cartesian product of any one of its horizontals with any one of its verticals. Just like for topological Markov shifts, the toral automorphism respects this structure: namely, for p $\in R_i \cap \phi^{-1} R_j$ the following holds:

$$\phi v_i(\mathrm{p}) \subset v_j(\phi \mathrm{p}),$$

$$\phi h_i(\mathrm{p}) \supset h_j(\phi \mathrm{p}).$$

We shall incorporate we have just described in a comprehensive theory.

4. More on Abstract Dynamical Systems.

DEFINITION 4.1. A dynamical system (X, ϕ) is said to be *irreducible* if for every pair of open sets U, V there exists $n \geq 0$ such that $\phi^n U \cap V \neq \emptyset$.

Another concept we need is the following

DEFINITION 4.2. A point p is said to be *bilaterally transitive* if the forward orbit $\{\phi^n \mathrm{p} | \; n \geq 0\}$ and the backward orbit $\{\phi^n \mathrm{p} | \; n < 0\}$ are both dense in X.

REMARK.. A symbolic sequence in a topological Markov shift is bilaterally transitive if every admissible block appears in both directions and infinitely often.

We use the notation $BLT(A)$ to denote the subset of bilaterally transitive points in $A \subseteq X$.

In an irreducible system the bilaterally transitive points turn out to be everywhere dense. To prove this, we recall the following theorem of point set topology. The theorem is more general, but can be slightly simplified in the case where the space X is a compact metric space.

BAIRE CATEGORY THEOREM 4.3. *Let $\{U_n\}$ be a countable collection of open dense subsets of X. Then $\bigcap U_n$ is nonempty. In fact $\bigcap U_n$ is dense in X. Equivalently, a compact metric space in not the union of a countable collection of nowhere dense sets.*

PROOF. Choose inductively balls B_n such that $B_n \subset \overline{B_n} \subset U_n$, and $\overline{B_n} \subset B_{n-1}$. The first property is easily achieved in a metric space; the second because U_n is dense which implies that $B_{n-1} \cap U_n$ is a non-empty open set. The sequence $(\overline{B_n})_{n \in \mathbb{N}}$ has the finite intersection property: so by compactness $\bigcap \overline{B_n}$ is nonempty. But $\bigcap \overline{B_n} \subset \bigcap B_n \subset \bigcap U_n$. Thus the intersection $\bigcap U_n$ is not empty. It is also dense which is a consequence of replacing U_n in the above argument by $U_n \cap B$ and X by $\overline{B}$ where B is any ball. $\square$

PROPOSITION 4.4. *If (X, ϕ) is irreducible, then the set of bilaterally transitive points is dense in X.*

PROOF. Let $\{U_n\}$ be a countable basis for X. Since X is irreducible, $\bigcup_{k \geq 0} \phi^k U_n$, as well as $\bigcup_{k<0} \phi^k U_n$, is dense in X for each n. The set $BLT(X)$ of bilaterally transitive points can be expressed as

$$BLT(X) = \bigcap_n \Big(\bigcup_{k \geq 0} \phi^k U_n \cap \bigcup_{k<0} \phi^k U_n\Big).$$

We apply the Baire Category Theorem to get $BLT(X) \neq \emptyset$. But $BLT(X)$ contains the whole orbit of any of its points and the orbit of any of its points is dense in X. $\square$

The three examples we have discussed exhibit a certain property which is easily verified. It is concerns the divergence of orbits and is defined as follows.

DEFINITION 4.5. A homeomorphism ϕ is said to be *expansive* if there exists a real number $c > 0$ such that if $d(\phi^n \mathrm{p}, \phi^n \mathrm{q}) < c$ for all $n \in \mathbb{Z}$ then $\mathrm{p} = \mathrm{q}$.

In the theory of Markov partitions this property plays a key role. Representation of dynamical systems by shifts of finite type, while possible for certain non-expansive systems, really only seems natural for expansive ones.

Next we formalize a property of mappings between dynamical systems previously alluded to in connection with binary expansions.

DEFINITION 4.6. For two general dynamical systems (X, ϕ) and (Y, ψ) we call the second a *factor* of the first and the first and *extension* of the second, if there exists a map π of X into Y, which we call a *factor map*, such that

(i) $\psi\pi = \pi\phi$,
(ii) π is continuous,
(iii) π is onto.

Furthermore, we say π is a *finite* factor map or that it is *bounded-to-one* if

(iv) there is a bound on the number of pre-images;

and *essentially*[4] *one-to-one* if

(v) every doubly transitive point has a unique pre-image.

$$\begin{tikzcd} X \arrow[r, "\phi"] \arrow[d, "\pi"'] & X \arrow[d, "\pi"] \\ Y \arrow[r, "\psi"'] & Y \end{tikzcd}$$

FIGURE 4.I. Commutative diagram illustrating a factor map

We remark that a topological conjugacy[5] is a finite factor map where the bound on the number of pre-images is one in condition 1.2(iv). As we shall see, the seemingly slight weakening of the chains of topological conjugacy, which is what the definition of an essentially one-to-one finite factor map is meant to do, allows the necessary freedom to get symbolic representations for smooth dynamical systems.

PROPOSITION 4.7. *Let π be a factor map of (X, ϕ) and (Y, ψ) : i.e. π satisfies properties (i), (ii), and (iii) of Definition 4.6. If (X, ϕ) is irreducible, then so is (Y, ψ); and $Y = \overline{BLT(Y)}$.*

PROOF. Let U, V be non-empty open subsets of Y. By properties (i) and (iii) of factor maps, $\pi^{-1}U, \pi^{-1}V$ are also non-empty and open. Since (X, ϕ) is irreducible,there exists $n > 0$ such that $\phi^n(\pi^{-1}U) \cap \pi^{-1}V \neq \emptyset$. By 4.6(i),

$$\emptyset \neq \pi[\phi^n(\pi^{-1}U) \cap \pi^{-1}V] = \pi[\pi^{-1}(\phi^n U \cap V)] = \psi^n U \cap V.$$

[4] This term is used because in irreducible systems the non-doubly transitive points are negligible in both the sense of category and measure.

[5] Ther term derives from the group theory notion of conjugate elements and its usage is standard in the subject. In the sense we are using it, better would have been *homomorphism* for *factor map* and *isomorphism* for *topological conjugacy*. These are the terms which denote the property of preserving structure.

Thus (Y, ψ) is irreducible and from Proposition 4.4 follows that $Y = \overline{BLT(Y)}$. $\square$

PROPOSITION 4.8. *Let (X, ϕ) be irreducible and π an essentially one-to-one factor map of (X, ϕ) onto (Y, ψ): i.e. π satisfies (i),(ii),(iii), and (v) of Definition 4.6. Then π maps $BLT(U)$ homeomorphically onto $BLT(\pi(U))$ for any open subset U of X.*

PROOF. From the properties of π, if the forward orbit of x hits every nonempty open subset of X, then the forward orbit of $\pi(x)$ hits every nonempty open subset of Y. Thus $\pi BLT(U) \subset BLT(\pi(U))$.

We have that π is a continuous one-to-one map of $BLT(U)$ into $BLT(\pi(U))$. We prove next that its inverse is continuous also. The proof is a standard compactness argument which goes as follows. Suppose $y_n \to y$, where $y_n, y \in BLT(\pi(U))$. We shall prove that $\pi^{-1}y_n \to \pi^{-1}y$. By compactness the sequence $(\pi^{-1}y_n)_{n\in\mathbb{N}}$ has limit points in X. Let x be any one of these limit points. By continuity $\pi x = y$. But the pre-image of y is unique: so the sequence $(\pi^{-1}y_n)_{n\in\mathbb{N}}$, having only one limit point, has a limit which is $\pi^{-1}y$.

Now let $x \in \pi^{-1}BLT(\pi(U)) \subset U$ and let V be any non-empty open subset of X. Choose $v \in BLT(V)$. Then by what we have already shown $\pi(v) \in BLT(\pi(V))$. Then there exists a sequence of positive integers k_n such that $\psi^{k_n}\pi(x) \to \pi(v)$. Thus

$$\phi^{k_n}(x) = \pi^{-1}\psi^{k_n}\pi(x) \to \pi^{-1}\pi(v) = v.$$

Thus $x \in BLT(U)$. We have therefore established $\pi^{-1}BLT(\pi(U)) \subset BLT(U)$: in other words, $BLT(\pi(U)) \subset \pi BLT(U)$. $\square$

PROPOSITION 4.9. *Under the hypothesis of Proposition 4.8, if U is an open subset of X, then $BLT(\pi U) = BLT[\pi U]^o$ and $\pi\overline{U} = \overline{[\pi U]^o}$.*

PROOF. Let $y \in BLT(\pi U)$. Then the unique pre-image of y lies in $BLT(U)$, and πy is not therefore in the closed set $\pi(X - U)$. Hence, $\pi y \in Y - \pi(X - U) \subset [\pi U]^o$. Therefore, $BLT(\pi U) = BLT[\pi U]^o$.

From the continuity properties of π, Proposition 4.8, and what was just proven, we get the following string of equalities: $\pi(\overline{U}) = \pi(\overline{BLT(U)}) = \overline{BLT(\pi U)} = \overline{BLT[\pi U]^o} =$ $\square$

Exercises.

4.1 We call a directed graph G *irreducible* if given any pair of nodes i, j there is a directed path from i to j. Show that if G is irreducible as a graph, then the dynamical system (Σ_G, σ) is irreducible. Conversely, show that if the dynamical system (Σ_G, σ) is irreducible then there is an irreducible sub-graph G' such that $\Sigma_G = \Sigma_{G'}$.

4.2 Given any topological Markov shift system (Σ_G) there exists irreducible sub-graphs $G_1, \ldots, G_M$ such that

$$\Sigma_G = \Sigma_{G_1} \cup \cdots \cup \Sigma_{G_M} \text{(disjoint)}.$$

5. Topological Partitions.

DEFINITION 5.1. We call a finite family of sets $\mathcal{R} = \{R_0, R_1, \ldots, R_{N-1}\}$ a *topological partition* for a compact metric space X if:

(1) Each R_i is open[6]
(2) $R_i \cap R_i = \emptyset, i \neq j$
(3) $X = \overline{R_0} \cup \overline{R_1} \cup \cdots \cup \overline{R_{N-1}}.$;

Remark. For open sets $U, V, \overline{U} \cap V \neq \emptyset \Rightarrow U \cap V \neq \emptyset$. So for members of a topological partition we get the following string of implications: $R_i \cap R_j = \emptyset \Rightarrow \overline{R_i} \cap R_j = \emptyset \Rightarrow \overline{R_i}^o \cap R_j = \emptyset \Rightarrow \overline{R_i}^o \cap \overline{R_j} = \emptyset \Rightarrow \overline{R_i}^o \cap \overline{R_j}^o = \emptyset$. Thus $\overline{R_i}^o \cap \overline{R_j}^o = \emptyset$ for $i \neq j$.

DEFINITION 5.2. Given two topological partitions $\mathcal{R} = \{R_0, R_1, \ldots, R_{N-1}\}$ and $\mathcal{S} = \{S_0, S_1, \ldots, S_{M-1}\}$ we define their *common topological refinement* $\mathcal{R} \vee \mathcal{S}$ as

$$\mathcal{R} \vee \mathcal{S} = \{R_i \cap S_j : R_i \in \mathcal{R}, S_j \in \mathcal{S}\}.$$

PROPOSITION 5.3. *The common topological refinement of two topological partitions is a topological partition.*

PROOF. Let $\mathcal{R}$ and $\mathcal{S}$ be the two partitions in question. First of all, it is clear that the elements of $\mathcal{R} \vee \mathcal{S}$ are disjoint. We show that the closure of elements of $\mathcal{R} \vee \mathcal{S}$ cover X. Let $\mathrm{p} \in X$. We have that $\mathrm{p} \in \overline{R_i}$ for some i. Thus there exists a sequence of points $\mathrm{p}_n \in R_i$ such that $d(\mathrm{p}_n, \mathrm{p}) < 1/n$. Since $\mathcal{S}$ is a topological partition, for each n there exists $S_{j_n} \in \mathcal{S}$ such that $\mathrm{p}_n \in \overline{S_{j_n}}$. Since $\mathcal{S}$ is finite, there exists an index j such that $j_n = j$ for an infinite number of n so that we can assume that the p_n were chosen in the first place such that each $j_n = j$. Since $\mathrm{p}_n \in R_i \cap \overline{S_j}$ we can choose a sequence of points $\mathrm{q}_{m,n} \in R_i \cap S_j$ such that $d(\mathrm{q}_{m,n}, \mathrm{p}_n) < 1/m$. Thus $d(\mathrm{q}_{n,n}, \mathrm{p}) < 2/n$ whence $\mathrm{q}_{n,n} \to \mathrm{p}$ as $n \to \infty$. Therefore $\mathrm{p} \in \overline{R_i \cap S_j}$.
□

PROPOSITION 5.4. *For dynamical system (X, ϕ) with topological partition $\mathcal{R}$ of X, the set $\phi^n \mathcal{R}$ defined by $\phi^n \mathcal{R} = \{\phi^n R_1, \ldots, \phi^n R_{N-1}\}$ is again a topological partition.*

PROOF. This is an immediate consequence of the following: (1)the image of a union is the union of images for any map; (2) a homeomorphism commutes with the operation of taking closures; (3)the image of an intersection is the intersection of images for a one-one map. □

From Proposition 5.3 and 5.4 we have that for $m \leq n$, $\bigvee_m^n \phi^k \mathcal{R} = \phi^m \mathcal{R} \vee \phi^{m-1} \mathcal{R} \vee \cdots \vee \phi^n \mathcal{R}$ is again a topological partition. We shall use the notation

$$\mathcal{R}^{(n)} \equiv \bigvee_{k=0}^{n-1} \phi^{-k} \mathcal{R}.$$

[6]Previous authors have taken these sets to be closed sets with the property that each is the closure of its interior. The present variation is slightly more general, just enough to make some notation and certain arguments simpler. In fact, an important example is presented in section 9 of a partition whose elements are not the interiors of their closures.

Thus $\mathcal{R}^{(2)} = \mathcal{R} \vee \phi^{-1}\mathcal{R} = \{R_i \cap \phi^{-1}R_j : R_i, R_j \in \mathcal{R}\}$. Observe that $(\mathcal{R}^{(2)})^{(2)} = \mathcal{R}^{(3)}$, or more generally $(\mathcal{R}^{(n)})^{(m)} = \mathcal{R}^{(n+m-1)}$.

The collection $\bigcup \phi^n \mathcal{R} : n \in \mathbb{Z}$ is a collection of open dense sets to which we can apply the Baire theorem, but due to its special nature we can achieve a slightly stronger result with the same sort of proof.

PROPOSITION 5.5. *Let $\mathcal{R}$ be a topological partition for dynamical system (X, ϕ). For every* p $\in X$ *there exists a sequence $(R_{s_k})_{k\in\mathbb{Z}}$ of sets in $\mathcal{R}$ such that* p $\in \bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^{n} \phi^{-k} R_{s_k}}$.

PROOF. Since $\bigvee_m^n \phi^k \mathcal{R},\ m \le n$, is a topological partition, there is a set in it whose closure contains p, say $\overline{\bigcap_m^n \phi^{-k} R_{s_k}}$. We next show that in the refinement $\bigvee_{m-1}^{n+1} \phi^k \mathcal{R}$ the elements of the form $\bigcap_{m-1}^{n+1} \phi^{-k} R_{t_k}$ where $t_k = s_k$ for $m \le k \le n$ comprise a subfamily which is a topological partition of $\overline{\bigcap_m^n \phi^{-k} R_{s_k}}$. Because $\bigvee_{m-1}^{n+1} \phi^k \mathcal{R}$ satisfies 5.1(1) and (2), so does any subfamily. Condition 5.1(3) is a consequence of

$$\bigcup_{\substack{0 \le t_{m-1} \le N-1 \\ 0 \le t_{n+1} \le N-1 \\ t_k = s_k,\ m \le k \le n}} \bigcap_{m-1}^{n+1} \phi^{-k} R_{t_k} = \bigcap_m^n \phi^{-k} R_{s_k}$$

and the fact that the closure of a union is the union of closures. Thus we can choose by induction the sets R_{s_k} as follows. Once having specified sets $R_{s_{-n}}, \dots, R_{s_n}$ such that p $\in \overline{\bigcap_{-n}^n \phi^{-k} R_{s_k}}$, we can find sets $R_{s_{-n-1}}$ and $R_{s_{n+1}}$ such that p $\in \overline{\bigcap_{-n-1}^{n+1} \phi^{-k} R_{s_k}}$. Hence there exists a sequence $(R_{s_k})_{k\in\mathbb{Z}}$ of sets in $\mathcal{R}$ such that p $\in \bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^n \phi^{-k} R_{s_k}}$. $\square$

REMARK.. With a slight modification of this proof somewhat more can be established: namely, a finite sequence of sets $R_{s_m}, R_{s_{m+1}}, \dots, R_{s_n}$ can be extended to a bi-infinite sequence $(R_{s_n})_{n\in\mathbb{Z}}$ such that if p $\in \overline{\bigcap_m^n \phi^{-k} R_{s_k}},\ m \le n$, then p $\in \bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^n \phi^{-k} R_{s_k}}$. We can even go further and make the same claim about extending a one-sided infinite sequence $R_{s_m}, R_{s_{m+1}}, \dots$ to a bi-infinite one.

DEFINITION 5.6. We define the *diameter* $d(\mathcal{R})$ of a partition $\mathcal{R}$ by

$$d(\mathcal{R}) = \max_{R_i \in \mathcal{R}} d(R_i)$$

where $d(R_i) \equiv \sup_{x,y \in R_i} d(x, y)$.

DEFINITION 5.7. We call a topological partition a *generator* for a dynamical system (X, ϕ) if $\lim_{n\to\infty} d\left(\bigvee_{-n}^n \phi^k \mathcal{R}\right) = 0$.

If $\mathcal{R}$ is a generator then clearly $\lim_{n\to\infty} d\left(\bigcap_{-n}^n \phi^{-k} R_{s_k}\right) = 0$ for any sequence of symbols $(s_i)_{i\in\mathbb{Z}} \in \{0, \dots, N-1\}^{\mathbb{Z}}$. The converse is also true (see exercise 5.1). In addition $d\left(\bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^n \phi^{-k} R_{s_k}}\right) = 0$. Hence in Proposition 5.5, if $\mathcal{R}$ is a generator and p $\in \bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^n \phi^{-k} R_{s_k}}$, then p $= \bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^n \phi^{-k} R_{s_k}}$.

The following proposition gives sufficient conditions on a topological partition in terms of its diameter for it to be a generator.

PROPOSITION 5.8.. *Let (X, ϕ) be expansive and $\mathcal{R}$ be a topological partition such that $d(\mathcal{R}) < c$ where c is the expansive constant. Then $\mathcal{R}$ is a generator;*

PROOF. The set $\bigcap_{-\infty}^{\infty} \phi^{-k}\overline{R_{s_k}}$ contains at most one point and thus has zero diameter: for if there exists $\mathrm{p}, \mathrm{q} \in \bigcap_{-\infty}^{\infty} \phi^{-k}\overline{R_{s_k}}$, then $d(\phi^n \mathrm{p}, \phi^n \mathrm{q}) < c$ for $n \in \mathbb{Z}$ implying $\mathrm{p} = \mathrm{q}$. Since $\bigcap_{-n}^{n} \phi^{-k}R_{s_k} \subset \bigcap_{-n}^{n} \phi^{-k}\overline{R_{s_k}}, d\left(\lim_{n\to\infty} \bigcap_{-n}^{n} \phi^{-k}R_{s_k}\right) = d\left(\bigcap_{-\infty}^{\infty} \phi^{-k}\overline{R_{s_k}}\right) = 0$. From Exercise 5.1 we get that $\mathcal{R}$ is a generator. $\square$

REMARKS. Generally we merely have the inclusion relation

$$\bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^{n} \phi^{-k}R_{s_k}} \subset \bigcap_{-\infty}^{\infty} \phi^{-k}\overline{R_{s_k}} \tag{5.9}$$

but not equality. However, when the sets of the partition are small enough– namely, when the hypothesis of Proposition 5.8 is satisfied– we do have equality: that is, if $\bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^{n} \phi^{-k}R_{s_k}} \neq \emptyset$ then $\bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^{n} \phi^{-k}R_{s_k}} = \bigcap_{-\infty}^{\infty} \phi^{-k}\overline{R_{s_k}}$.

From the inclusion relation (5.9) we see that if $\pi(s) = \mathrm{p}$, then $\mathrm{p} \in \overline{R_{s_0}}$. Thus if x belongs only to $\overline{R_i}$, then $s_0 = i$. In particular, by the remark following Definition 5.1, if $\mathrm{p} \in R_i$ or $\mathrm{p} \in \overline{R_i}^o$, then $s_0 = i$.. In addition, if there exists sequences s, t such that $\pi(s) = \pi(t) = \mathrm{p}$ and $s_0 = i \neq j = t_0$, then $\mathrm{p} \in \overline{R_i} \cap \overline{R_j}$, $i \neq j$; and conversely. In which case $\mathrm{p} \in (\overline{R_i} - \overline{R_i}^o) \cap (\overline{R_j} - \overline{R_j}^o) = \partial R_i \cap \partial R_j$: *i.e.* p belongs to the boundary of partition elements.

Let $\mathcal{R} = \{R_1, \dots, R_N\}$ be a generator for a dynamical system (X, ϕ) Let Σ be the subset of the full N-shift defined by

$$\Sigma \equiv \{s = (\dots, s_n, \dots) : \bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^{n} \phi^{-k}R_{s_k}} \neq \emptyset\}. \tag{5.10}$$

Because the topological partition $\mathcal{R}$ is a generator, the non-empty infinite intersection $\bigcap_{n=0}^{\infty} \overline{\bigcap_{k=-n}^{n} \phi^{-k}R_{s_k}}$ consists of a single point. Therefore, we can define a map $\pi : \Sigma \longrightarrow X$ by

$$\pi(s) = \bigcap_{n=0}^{\infty} \overline{\phi^n R_{s_{-n}} \cap \phi^{n-1} R_{s_{-n+1}} \cap \dots \cap \phi^{-n} R_{s_n}}. \tag{5.11}$$

$$\begin{tikzcd} \Sigma_G \arrow[r, "\sigma"] \arrow[d, "\pi"'] & \Sigma_G \arrow[d, "\pi"] \\ X \arrow[r, "\phi"'] & X \end{tikzcd}$$

FIGURE 5.I. Commutative diagram for symbolic representation

PROPOSITION 5.12.. *Let the dynamical system (X, ϕ) have a topological partition $\mathcal{R}$ which is a generator. Then Σ as defined by (5.10) is a closed shift-invariant subset of Σ_N and the map π given by (5.11) is a factor map of the dynamical system (Σ, σ) onto (X, ϕ) — — i.e., π satisfies the following items of Definition 4.6:*

(i) $\sigma\pi = \pi\phi$,
(ii) π *is continuous,*
(iii) π *is onto.*

PROOF. To prove Σ is closed we must show that if $s = (\ldots, s_k, \ldots) \in \overline{\Sigma}$ then

$$\bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^{n} \phi^{-k} R_{s_k}} \neq \emptyset, \tag{5.13}$$

which then implies that $s \in \Sigma$. For each $n > 1$ there is a sequence $t = (\ldots, t_k, \ldots) \in \Sigma$ such that $d(t, s) < 1/n$. This means that $t_k = s_k, -n \leq k \leq n$ so that

$$\overline{\bigcap_{-n}^{n} \phi^{-k} R_{t_k}} = \overline{\bigcap_{-n}^{n} \phi^{-k} R_{s_k}}.$$

Because $t \in \Sigma$, this set is non-empty. Since this is so for arbitrary n and these sets form a decreasing sequence of non-empty closed sets, applying compactness we get 5.13.

To prove Σ is $\sigma-$invariant we must show that if $s \in \Sigma$ then $\sigma s \in \Sigma$: in other words, for $n \geq 0$, if $\bigcap_{-n}^{n} \phi^{-k} R_{s_k} \neq \emptyset$ then $\bigcap_{-n}^{n} \phi^{-k} R_{s_{k+1}} \neq \emptyset$. This follows from using the distributive property of ϕ^{-1} with respect to intersections and reindexing: *i.e.*

$$\phi^{-1}\left(\bigcap_{-n}^{n} \phi^{-k} R_{s_{k+1}}\right) = \bigcap_{-n+1}^{n+1} \phi^{-k} R_{s_k} \supset \bigcap_{-n+1}^{n-1} \phi^{-k} R_{s_k}.$$

We now turn our attention to the properties of π.
(i) π satisfies $\pi\sigma = \phi\pi$. This follows from reindexing after applying the property that a homeomorphism commutes with the closure operation and preserves intersections: to wit,

$$\begin{aligned}
\phi\pi(s) &= \phi \bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^{n} \phi^{-k} R_{s_k}} \\
&= \bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^{n} \phi^{-k+1} R_{s_k}} \\
&= \bigcap_{n=0}^{\infty} \overline{\bigcap_{-n-1}^{n-1} \phi^{-k} R_{s_{k+1}}} \\
&= \bigcap_{n=0}^{\infty} \overline{\bigcap_{-n+1}^{n-1} \phi^{-k} R_{s_{k+1}}} = \pi(\sigma s).
\end{aligned}$$

(ii) π is continuous. From the generating property of $\mathcal{R}$, given $\epsilon > 0$ there is a positive integer n such that $d(\overline{\bigcap_{-n}^{n} \phi^{-k} R_{s_k}}) < \epsilon$. Thus, for $s, t \in \Sigma_G$, there is a $\delta > 0$, namely $\delta = 1/(n+1)$, such that if $d(s,t) < \delta$ then $\pi(s), \pi(t) \in \overline{\bigcap_{-n}^{n} \phi^{-k} R_{s_k}}$.

(iii) π is onto. This follows immediately from Proposition 5.5.

$\square$

Exercises.

5.1 Let $\mathcal{R}$ be a topological partition for a dynamical system (X, ϕ). Prove that if $\lim_{n\to\infty} d\left(\bigcap_{-n}^{n} \phi^{-k} R_{s_k}\right) = 0$ for any sequence of symbols $(s_i)_{i\in\mathbb{Z}} \in \{0, \ldots, N-1\}^{\mathbb{Z}}$ then $\lim_{n\to\infty} d\left(\bigvee_{-n}^{n} \phi^{k} \mathcal{R}\right) = 0$.

6. Markov partitions and symbolic extensions.

DEFINITION 6.1. We say that a topological partition $\mathcal{R}$ for a dynamical system (X, ϕ) satisfies *the n-fold intersection property* for a positive integer $n \geq 3$, if $R_{s_k} \cap \phi^{-1} R_{s_{k+1}} \neq \emptyset,\ 1 \leq k \leq n-1 \Rightarrow \bigcap_{k=1}^{n} \phi^{-k} R_{s_k} \neq \emptyset$. Furthermore, we call a topological partition *Markov* if it satisfies the n-fold intersection property for all $n \geq 3$.

REMARK.. In section 3.1 and before the term "Markov" topological generator was defined, we considered the partition $\mathcal{C} = \{C_a : a \in A\}$ consisting of the elementary cylinder sets $C_a = \{s \in \Sigma_G : s_0 = a\}$ for a dynamical system (Σ_G, σ) where Σ_G is a shift of finite type base on an alphabet A. As one might have guessed this partition is the prototype of a topological Markov generator.

PROPOSITION 6.2. *If $\mathcal{R}$ is a Markov partition, then so is $\bigvee_{m}^{n} \phi^{k} \mathcal{R}$ for any $m \leq n$.*

PROOF. We leave the proof as an exercise. $\square$

If a topological partition $\mathcal{R}$ satisfies the n-fold intersection property, then it satisfies k-fold ones for all smaller k. To increase the order we shall utilize the following.

BOOTSTRAP LEMMA 6.3. *If $\mathcal{R}$ satisfies the 3-fold and $\mathcal{R}^{(2)}$ satisfies the n-fold intersection properties, $n \geq 3$, then $\mathcal{R}$ satisfies the $(n+1)$-fold intersection property.*

PROOF. Suppose $R_{i_k} \cap \phi^{-1} R_{i_{k+1}} \neq \emptyset,\ 1 \leq k \leq n$. Because $\mathcal{R}$ satisfies the 3-fold intersection property, we have

$$R_{i_k} \cap \phi^{-1} R_{i_{k+1}} \cap \phi^{-2} R_{i_{k+2}} \neq \emptyset,\ 1 \leq k \leq n-1.$$

In other words,

$$(R_{i_k} \cap \phi^{-1} R_{i_{k+1}}) \cap \phi^{-1}(R_{i_{k+1}} \cap \phi^{-1} R_{i_{k+2}}) \neq \emptyset,\ 1 \leq k \leq n-1.$$

Because $\mathcal{R}^{(2)}$ satisfies the n-fold intersection property, we obtain

$$\bigcap_{1}^{n+1} \phi^{-k} R_{i_k} = \bigcap_{1}^{n} \phi^{-k}(R_{i_k} \cap \phi^{-1} R_{i_{k+1}}) \neq \emptyset.$$

□

Suppose a dynamical system (X, ϕ) has a Markov generator $\mathcal{R} = \{R_0, \dots, R_{N-1}\}$. We define an associated topological Markov shift given by the directed graph G whose vertices are labelled by $\mathcal{A} = \{0, 1, \dots, N-1\}$ and in which the i-the vertex is connected to the j-th, $i \to j$, iff $R_i \cap \phi^{-1} R_j \neq \emptyset$. So by definition of the Markov shift associated with a transition matrix of a directed graph,

$$\Sigma_G = \{s = (s_n)_{n \in \mathbb{Z}} : R_{s_{n-1}} \cap \phi^{-1} R_{s_n} \neq \emptyset, s_n \in \mathcal{A}, n \in \mathbb{Z}\}. \tag{6.4}$$

This set coincides with the subsystem defined by 5.10 which is easily seen as follows. On one hand, for $s \in \Sigma_G$, each of the closed sets

$$\{ \overline{\bigcap_{k=-n}^{n} \phi^{-k} R_{s_k}} \mid n = 1, 2 \dots \}$$

for any $n \geq 0$ is nonempty since the finite intersection under the closure sign is nonempty due to the Markov property. For increasing n these closed intersections form a decreasing sequence of non-empty sets; and therefore by compactness $\bigcap_{n=0}^{\infty} \overline{\bigcap_{k=-n}^{n} \phi^{-k} R_{s_k}} \neq \emptyset$. On the other hand, if $\bigcap_{n=0}^{\infty} \overline{\bigcap_{k=-n}^{n} \phi^{-k} R_{s_k}} \neq \emptyset$, then each finite intersection under the closure sign is non-empty which in turn implies that each pair of intersections $R_{s_k} \cap \phi^{-1} R_{s_k} \neq \emptyset$, for arbitrary $k \in \mathbb{Z}$.

Main Theorem.

THEOREM 6.5. *Suppose the dynamical system (X, ϕ) is expansive and has a Markov generator $\mathcal{R} = \{R_0, \dots, R_{N-1}\}$. Then the map π, as defined by (5.11), is an essentially one-to-one finite factor map of the shift of finite type Σ_G, as defined by (6.1), onto X. Furthermore, If (X, ϕ) is irreducible, then so is (Σ_G, σ).*

PROOF. We must establish (i) - (v) in Definition 4.6. That π is a factor map–namely, it satisfies items (i), (ii), and (iii)–is the content of theorem 5.12.

In order to establish *(iv)*–namely, a bound on the number of pre-images under π–we introduce the following concept.

DEFINITION 6.6. A map π from Σ_G to X said to have a *diamond* if there are two sequences $s, t \in \Sigma_G$ for which $\pi(s) = \pi(t)$ and for which there exist indices $k < l < m$ such that $s_k = t_k, s_l \neq t_l, s_m = t_m$.

LEMMA 6.7. *If the number of pre-images of a point is more than N^2, then π has a diamond.*

PROOF. We apply the familiar "pigeon hole" argument. Let $s^{(1)}, \dots s^{(N^2+1)}$ be N^2+1 different sequences which map to the same point. Since the sequences are distinct, there are a pair of indices k, m such that the allowable blocks $s^{(1)}_{[k,m]}, \dots s^{(N^2+1)}_{[k,m]}$

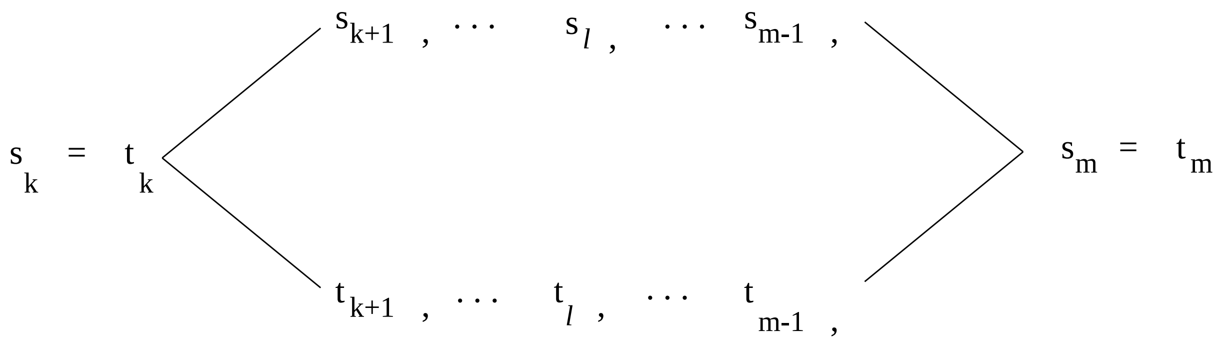


FIGURE 6.I. A diamond

are distinct. There are N^2 distinct choices of pairs of symbols $(s_k^{(i)}, s_m^{(i)})$: so by the "pigeon hole principle" there must be two allowable blocks $s_{[k,m]}^{(i)}, s_{[k,m]}^{(j)}$, such that $(s_k^{(i)}, s_m^{(i)}) = (s_k^{(j)}, s_m^{(j)})$. But, since the blocks are different, there is an index l such that $s_l^{(i)} \neq s_l^{(j)}$. Thus the two sequences $s^{(i)}, s^{(j)}$ map to the same point, agree at indices k, m, but differ at l which is between k, m, which means there is a diamond. □

LEMMA 6.8. *If there exists a bilaterally transitive point with two pre-images, then π has a diamond.*

PROOF. Let a BLT p have two pre-images. As we have indicated in the remark following 5.8, there are two sets $R_a, R_b \in \mathcal{R}, a \neq b$, such that p $\in \overline{R_a} \cap \overline{R_b}$. For each $n > 0$, the family of sets

$$\{\overline{\phi^n R_{s_{-n}} \cap \cdots \cap R_{s_0} \cap \ldots \phi^{-n} R_{s_n}} : s \in \Sigma_G \quad \text{where} \quad s_0 = a\}$$

covers $\overline{R_a}$; and the family

$$\{\overline{\phi^n R_{t_{-n}} \cap \cdots \cap R_{t_0} \cap \ldots \phi^{-n} R_{t_n}} : t \in \Sigma_G \quad \text{where} \quad t_0 = b\}$$

covers $\overline{R_b}$. Thus, by compactness, there exists $s, t \in \Sigma_G$ with $s_0 = a, t_0 = b$ such that

$$\mathrm{p} \in \bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^{n} \phi^{-k} R_{s_k}}$$

and

$$\mathrm{p} \in \bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^{n} \phi^{-k} R_{t_k}}$$

Since p is bilaterally transitive and R_0 is open, $\phi^n \mathrm{p} \in R_0$ for some positive n and $\phi^m \mathrm{p} \in R_0$ for some negative m. Thus by the remark following (5.8), $s_m = t_m = 0, s_0 = a \neq b = t_0, s_n = t_n = 0$, which is a diamond for π. □

LEMMA 6.9. *If $d(\mathcal{R}) < c/2$, then π has no diamonds.*

PROOF. Since $\pi\sigma = \phi\pi$, we can asume without loss of generality that $k = 1$ in the definition of diamond. Assume that $\mathrm{p} = \pi(s) = \pi(t)$ where

$$s = (\dots, s_{-2}, a, b_0, b_1, \dots, b_{m-1}, d, s_{m+1}, \dots),$$

$$t = (\dots, t_{-2}, a, c_0, c_1, \dots, c_{m-1}, d, t_{m+1}, \dots).$$

We must show that $b_l = c_l$ for $0 \le l \le m-1$. Because $[a, b_0, b_1, \dots, b_{m-1}, d]$ is an allowable block in Σ_G,

$$\phi R_a \cap R_{b_0} \cap \phi^{-1} R_{b_1} \cap \dots \cap \phi^{-m+1} R_{m-1} \cap \phi^{-m} R_d \neq \emptyset.$$

Choose a point q in this open set. Because π is onto, there is a sequence

$$u = (\dots, u_{-2}, a, b_0, b_1, \dots, b_{m-1}, d, u_{m+1}, \dots) \in \Sigma_G$$

such that $\pi(u) = \mathrm{q}$. Also since $[a, c_0, c_1, \dots, c_{m-1}, d]$ is an allowable block and Σ_G is a shift of finite type, there is a sequence $v \in \Sigma_G$ such that

$$v = (\dots, u_{-2}, a, c_0, c_1, \dots, c_{m-1}, d, u_{m+1}, \dots).$$

Thus

$$\mathrm{r} \equiv \pi(v) \in \overline{\phi R_a \cap R_{c_0} \cap \phi^{-1} R_{c_1} \cap \dots \cap \phi^{-m+1} R_{m-1} \cap \phi^{-m} R_d}.$$

From $d(R_i) < c/2$ and $\phi^l(x) \in \overline{R_{b_l}} \cap \overline{R_{c_l}}$ for $0 \le l \le m-1$, we conclude by the triangle inequality that $d(\phi^l \mathrm{q}, \phi^l \mathrm{r}) < c$. Furthermore, $d(\phi^n \mathrm{q}, \phi^n \mathrm{r}) < c/2$ for $n < 0$ and $n > m-1$. The expansive property then implies that $\mathrm{q} = \mathrm{r}$. Thus $R_{b_l} \cap \overline{R_{c_l}} \neq \emptyset$ which implies that $R_{b_l} \cap R_{c_l} \neq \emptyset$. However, elements of $\mathcal{R}$ are pairwise disjoint: so $b_l = c_l$. $\square$

(iv) There is a bound on the number of pre-images of π.
(v) A BLT point has a unique pre-image.

Because $\mathcal{R}$ is a generator, n can be chosen so that $d\left(\bigvee_{-n}^{n} \phi^k \mathcal{R}\right) < c/2$. By Proposition 6.2, $\bigvee_{-n}^{n} \phi^k \mathcal{R}$ is again a topological Markov partition. For this partition the associated shift of finite type of (6.4) is given by the higher edge graph $G^{(2n+1)}$. Let $\pi^{(2n+1)}$ be the map of $\Sigma_{G^{(2n+1)}}$ onto X according to (5.11). It has no diamonds: so by Lemma 6.7 a point has at most $N^{2(2n+1)}$ pre-images, and by Lemma 6.8 a BLT point has only one. The original π satisfies $\pi = \pi^{(2n+1)} \psi \sigma^n$ where ψ a conjugacy of Σ_G onto $\Sigma_{G^{(2n+1)}}$. Thus we have that under π a point has at most $N^{2(2n+1)}$ pre-images, and a BLT point has a unique pre-image.

We defer the proof of irreducibility to Exercise 6.2 $\square$

Converse to the Main Theorem.
Recall that we introduced in 3.1 the partition $\mathcal{C} = \{C_i : i = 0, \dots, N-1\}$ consisting of the elementary cylinder sets $C_i = \{s \in \Sigma_G : s_0 = i\}$ for a dynamical system (Σ_G, σ) where Σ_G is a shift of finite type base on an alphabet $A = \{0, 1, \dots, N-1\}$. This partition is a topological Markov generator.

THEOREM 6.10. *Let (X,ϕ) be a dynamical system, (Σ_G,σ) an irreducible shift of finite type based on N symbols; and suppose there exists an essentially one-to-one factor map π from Σ_G to X. Then the partition $\mathcal{R}$ defined by $\mathcal{R} = \{R_i = \pi(C_i)^o : i = 0,\dots,N-1\}$ is a topological Markov generator.*

REMARK. Note we assume π is a factor map which has a unique inverse for each bilaterally transitive point, but no bound is assumed on the number of pre-images of arbitrary points: *i.e.*, π satisfies (i),(ii),(iii), and (v) of Definition 4.6 but not (iv). However, in Corollary 6.12 we shall show that (iv) follows from the others under the hypothesis of expansivity. However, as Exercise 6.3 shows property (v) is essential: we cannot obtain it from expansivity and (i) through (iv).

PROOF.
We must prove the following items

(1) Elements of $\mathcal{R}$ are disjoint:
(2) The closure of elements of $\mathcal{R}$ cover X.
(3) $\mathcal{R}$ is a generator.
(4) $\mathcal{R}$ satisfies the Markov property.

(1) Elements of $\mathcal{R}$ are disjoint: i.e., $R_i \cap R_j = \emptyset, i \neq j$.
The idea of the proof is to use bilaterally transitive points to overcome a difficulty: namely, maps in general do not enjoy the property that the image of an intersection is equal to the intersection of images, but one-to-one maps do. Suppose $R_i \cap R_j \neq \emptyset$ for $i \neq j$. Then, by Proposition 4.7 $BLT(R_i \cap R_j) \neq \emptyset$. By Propositions 4.8 and 4.9, π^{-1} maps $BLT(R_i)$ and $BLT(R_j)$ homeomorphically onto $BLT(C_i)$ and $BLT(C_j)$ respectively. Therefore π^{-1} maps $BLT(R_i \cap R_j) = BLT(R_i) \cap BLT(R_j)$ homeomorphically onto $BLT(C_i \cap C_j) = BLT(C_i) \cap BLT(C_j)$, which implies that $\emptyset \neq BLT(C_i \cap C_j) \subset C_i \cap C_j$, a contradiction.
(2) $X = \cup_{i=0}^{N-1} \overline{R_i}$.
$X = \pi(\Sigma_G) = \pi \cup_{i=0}^{N-1} C_i = \cup_{i=0}^{N-1} \pi(C_i) = \cup_{i=0}^{N-1} \overline{R_i}$, the last inequality following from Proposition 4.9.
For the next two items we need a lemma.

LEMMA 6.11. *Under the hypothesis of 6.10, $\pi(\bigcap_m^n \sigma^{-k} C_{s_k}) = \overline{\bigcap_m^n \phi^{-k} R_{s_k}}$ for $m < n$.*

PROOF. Once again we use the bilaterally transitive points to deal with images of intersections. We have the following string of equalities.

$$\pi(\bigcap_m^n \sigma^{-k} C_{s_k}) = \pi\left(\overline{BLT(\bigcap_m^n \sigma^{-k} C_{s_k})}\right) = \overline{\pi(\bigcap_m^n BLT(\sigma^{-k} C_{s_k}))}$$

which by injectivity of π and shift-invariance of bilateral transitive points

$$= \overline{\bigcap_m^n \pi(BLT(\sigma^{-k} C_{s_k}))} = \overline{\bigcap_m^n \pi \sigma^{-k} BLT((C_{s_k}))}$$

which by commutativity of π and Proposition4.8

$$= \overline{\bigcap_m^n \phi^{-k} \pi(BLT(C_{s_k}))} = \overline{\bigcap_m^n \phi^{-k}(BLT\pi(C_{s_k}))}$$

which by Proposition 4.9

$$
\begin{aligned}
&= \overline{\bigcap_m^n \phi^{-k}(BLT(R_{s_k}))} = \overline{BLT(\bigcap_m^n \phi^{-k}(R_{s_k}))} \\
&= \overline{\bigcap_m^n \phi^{-k}(R_{s_k})}.
\end{aligned}
$$

Q.E.D.

(3) $\mathcal{R}$ is a generator.

Because $\mathcal{C}$ is a generator, $d\left(\bigcap_{-n}^n \sigma^{-k} C_{s_k}\right) \to 0$. By Lemma 6.11 , $\pi(\bigcap_{-n}^n \sigma^{-k} C_{s_k}) = \overline{\bigcap_{-n}^n \phi^{-k} R_{s_k}}$. So, by continuity of π, we get

$$
d\left(\overline{\bigcap_{-n}^n \phi^{-k} R_{s_k}}\right) = d\left(\bigcap_{-n}^n \phi^{-k} R_{s_k}\right) \to 0.
$$

(4) $\mathcal{R}$ satisfies the Markov property.

Suppose $R_{s_i} \cap \phi^{-1} R_{s_{i+1}} \neq \emptyset, 1 \leq k \leq n-1$. By Lemma 6.11 we have $\pi[C_{s_i} \cap \sigma^{-1} C_{s_{i+1}}] = \overline{R_{s_i} \cap \phi^{-1} R_{s_{i+1}}} \neq \emptyset, 1 \leq k \leq n-1$. Thus $C_{s_i} \cap \sigma^{-1} C_{s_{i+1}} \neq \emptyset, 1 \leq k \leq n-1$. Since $\mathcal{C}$ satisfies the Markov property, $\bigcap_{k=1}^n \phi^{-k} C_{s_k} \neq \emptyset$ for all $n > 1$. So $\pi(\bigcap_{k=1}^n \phi^{-k} C_{s_k}) = \overline{\bigcap_{k=1}^n \phi^{-k} R_{s_k}} \neq \emptyset$ for all $n > 1$. Therefore, $\bigcap_{k=1}^n \phi^{-k} R_{s_k} \neq \emptyset$ for all $n > 1$. □

Corollary 6.12. *If in addition to the hypotheses of Theorem 4.18 the dynamical system (X, ϕ) is expansive, then π is finite.*

Proof. We derive 4.6(iv) from from the assumption that the domain of π is irreducible, π satisfies 4.6(i), (ii), (iii), and (v), and ϕ is expansive. This is an immediate consequence of Theorems 6.10 and 6.5. □

Exercises.

6.1 A topological partition $\mathcal{R}$ for a dynamical system (X, ϕ) is Markov if and only if

$$
\bigcap_{k=0}^n \phi^{-k} R_{s_k} \neq \emptyset, \bigcap_{k=-n}^0 \phi^{-k} R_{s_k} \neq \emptyset \Rightarrow \bigcap_{k=-n}^n \phi^{-k} R_{s_k} \neq \emptyset,
$$

for $n \geq 0$.

6.2 Under the hypothesis of Theorem 6.5 show that if (X, ϕ) is irreducible, then so is (Σ_G, σ).

6.3 Show that in Theorem 6.10, condition (v) cannot be replaced by condition (iv). Hint consider the following example. Let

$$
\mathrm{A} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix},
$$

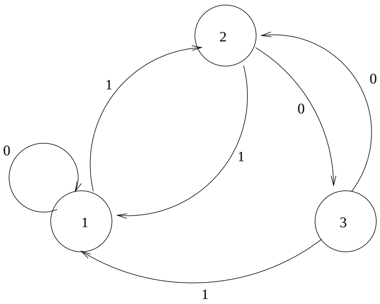


FIGURE 6.II. π indicated by edge-labels

and $\pi : X_{\mathrm{A}} \to X_2$ be defined by

$$\pi[1,1] = \pi[2,3] = \pi[3,2] = 1$$

$$\pi[1,2] = \pi[2,1] = \pi[3,1] = 0.$$

Show disjointness is violated in the image partition of the elementary cylinder sets.

6.3 Let π be a factor map from a topological Markov shift $(\Sigma_{\mathrm{A}}, \sigma)$ to an expansive dynamical system (X, ϕ) with expansive constant c and let the partition $\mathcal{R}$ defined by $\mathcal{R} = \{R_i = \pi(C_i)^o : i = 0, \ldots, N-1\}$ satisfy $d(\mathcal{R}) < c/2$. Show if π has a diamond, then there exists a point $\mathrm{p} \in X$ with a continuum of pre-images.

7. Product Structure.

The Markov property for a topological partition is an infinite set of conditions. It is the crucial one for obtaining a topological Markov shift representation of a dynamical system, but it could be difficult to verify. However there is another more useful criterion for getting it to which we now now turn our attention. It involves exchanging one set infinite set of conditions for another of a different sort which are more readily checkable. Once more we looked to our concrete systems as as guide. The sets of the partitions in examples 3.1 and 3.3 have a product structure whose behavior with respect to action of a mapping is intimately tied up with the Markov property.

A general notion of partition without regard to any other consideration is the following.

DEFINITION 7.1. A *partition* of a set R is defined to be a family $\mathcal{H} = \{h(\mathrm{p}) : \mathrm{p} \in R\}$ of subsets of R such that for $\mathrm{p}, \mathrm{q} \in R$

(1) $\mathrm{p} \in h(\mathrm{p})$
(2) $h(\mathrm{p}) \cap h(\mathrm{q}) \neq \emptyset \Rightarrow h(\mathrm{p}) = h(\mathrm{q})$.

DEFINITION 7.2. We call two partitions

$$\mathcal{H} = \{h(\mathrm{p}) : \mathrm{p} \in R\},$$
$$\mathcal{V} = \{v(\mathrm{p}) : \mathrm{p} \in R\}$$

of R *transverse* if, $h(\mathrm{p}) \cap v(\mathrm{q}) \neq \emptyset$ for every $\mathrm{p}, \mathrm{q} \in R$.

A set R with two transverse partitions $\mathcal{H}$, $\mathcal{V}$, can be viewed as having a product structure something like that of a rectangle which suggests the following graphic terminology: we shall refer to the elements $\mathcal{H}$ as *horizontals* and those of $\mathcal{V}$ as *verticals*. When we are dealing with elements R_i of a topological partition $\mathcal{R} = \{R_i : i = 1, \ldots, N\}$, each of having a pair of transverse partitions $\mathcal{H}_i, \mathcal{V}_i$, we refer to $h_i(\mathrm{p})$, as the *horizontal through* p *in* R_i and to $v_i(\mathrm{p})$ as the *vertical*.

Next we introduce notions concerned with the behavior of horizontals and verticals under the map associated with a dynamical system. We shall stick to the convention that under the action of a map verticals seem to contract and horizontals seem to expand. While we don't insist that the diameters of the images of these sets actually increase or decrease, this will generally be the case. In the literature one encounters the term *stable* set for what we call a vertical and *unstable* set for a horizontal.

DEFINITION 7.3. Suppose a dynamical system (X, ϕ) has a topological partition $\mathcal{R} = \{R_i\}$, each member of which has a pair of transverse partitions. We say *alignment* of verticals and horizontals are respectively *maintained by* by ϕ and ϕ^{-1} if for all i, j

(1) $\mathrm{p} \in R_i \cap \phi^{-1} R_j \Rightarrow R_j \cap \phi v_i(\mathrm{p}) \subset v_j(\phi \mathrm{p})$;
(2) $\mathrm{p} \in R_i \cap \phi R_j \Rightarrow R_j \cap \phi^{-1} h_i(\mathrm{p}) \subset h_j(\phi^{-1} \mathrm{p})$.

We actually require something stronger.

DEFINITION 7.4. In a dynamical system (X, ϕ) we say a topological partition $\mathcal{R} = \{R_i\}$ has *property M* if each set R_i has a pair of transverse partitions such that alignments of horizontals and verticals are maintained by ϕ and its inverse respectively in such a manner that the image of any vertical and the pre-image of any horizontal is contained in a unique element of $\mathcal{R}$. In other words 7.3 (1) and (2) are replaced by:

(1) $\mathrm{p} \in R_i \cap \phi^{-1} R_j \Rightarrow \phi v_i(\mathrm{p}) \subset v_j(\phi \mathrm{p})$,
(2) $\mathrm{p} \in R_i \cap \phi R_j \Rightarrow \phi^{-1} h_i(\mathrm{p}) \subset h_j(\phi^{-1} \mathrm{p})$.

We remark that with respect to horizontals 7.6(2) can be expressed alternatively as follows:

$$\mathrm{p} \in R_i \cap \phi^{-1} R_j \Rightarrow \phi h_i(\mathrm{p}) \supset h_j(\phi \mathrm{p}).$$

PROPOSITION 7.5. *If $\mathcal{R}$ has property M, then so does $\mathcal{R}^{(2)}$.*

PROOF. Let $R_i \cap \phi^{-1} R_j \neq \emptyset$ be a member of $\mathcal{R}^{(2)}$. Since a partition of a set induces one on a subset, the horizontals and verticals of R_i induce corresponding partitions of $R_i \cap \phi^{-1} R_j$: namely,

(1) $\mathcal{H}(R_i \cap \phi^{-1} R_j) = \{h_{ij}(\mathrm{p}) = h_i(\mathrm{p}) \cap \phi^{-1} R_j : \mathrm{p} \in R_i \cap \phi^{-1} R_j\}$,
(2) $\mathcal{V}(R_i \cap \phi^{-1} R_j) = \{v_{ij}(\mathrm{p}) = v_i(\mathrm{p}) : \mathrm{p} \in R_i \cap \phi^{-1} R_j\}$.

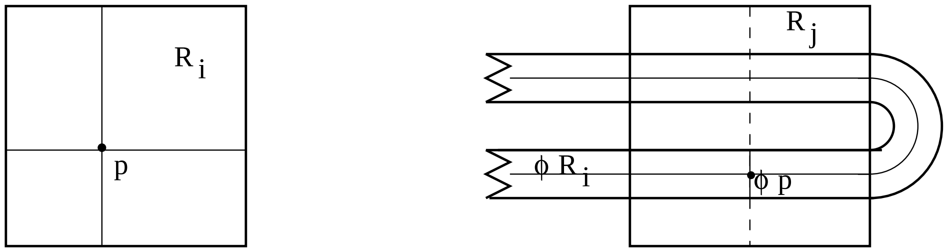


FIGURE 7.I. Property M

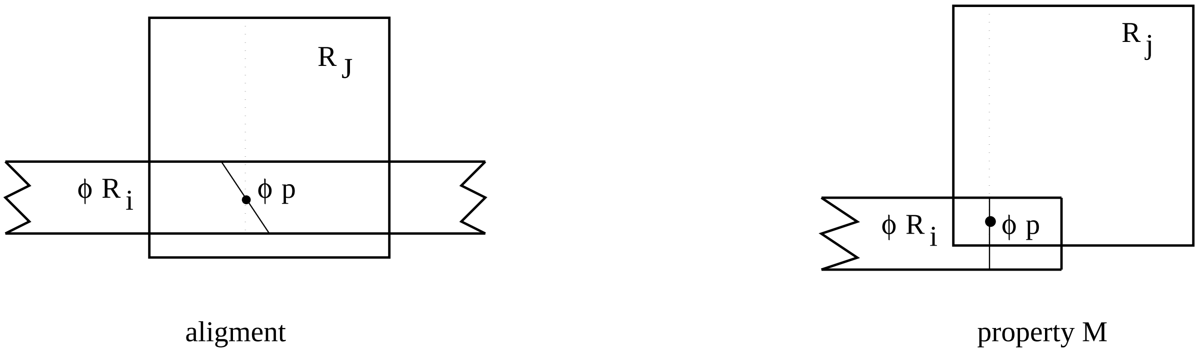


FIGURE 7.II. Violations

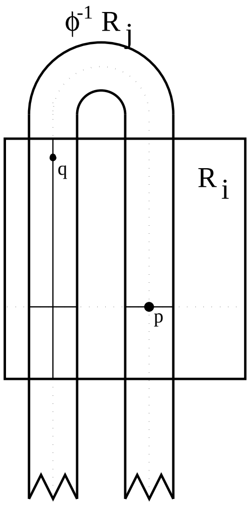


FIGURE 7.III. Property M on $\mathcal{R}^{(2)}$

First to verify that this pair of partitions is transverse, we observe that if p, q $\in$ $R_i \cap \phi^{-1} R_j$ then by 7.4(1)

$$v_i(\mathrm{q}) \subset \phi^{-1} v_j(\phi \mathrm{q}) \subset \phi^{-1} R_j.$$

From definition (1) we have

$$h_{ij}(\mathrm{p}) \cap v_{ij}(\mathrm{q}) = h_i(\mathrm{p}) \cap \phi^{-1}R_j \cap v_i(\mathrm{q}) = h_i(\mathrm{p}) \cap v_i(\mathrm{q}) \neq \emptyset.$$

Second, we show that ϕ and its inverse map verticals and horizontals so as to satisfy property M. Let p $\in R_i \cap \phi^{-1}R_j \cap \phi^{-1}(R_j \cap \phi^{-1}R_k)$. On one hand, it is immediate from definition that

$$\phi v_{ij}(\mathrm{p}) = \phi v_i(\mathrm{p}) \subset v_j(\phi\mathrm{p}) = v_{ij}(\phi\mathrm{p}).$$

On the other,

$$\phi h_{ij}(\mathrm{q}) = \phi(h_i(\mathrm{q}) \cap \phi^{-1}R_j) \supset h_j(\phi\mathrm{q}) \supset h_j(\phi\mathrm{q}) \cap \phi^{-1}R_k = h_{jk}(\phi\mathrm{q}).$$

□

COROLLARY 7.6. *If $\mathcal{R}$ has property M then so does $\mathcal{R}^{(n)}$ for $n = 1, 2, \dots$.*

PROOF. Repeated use Proposition 7.5 using the identity $(\mathcal{R}^{(n)})^{(2)} = \mathcal{R}^{(n+1)}$. □

PROPOSITION 7.7. *For a dynamical system (X, ϕ) if a topological partition $\mathcal{R}$ has property M, then $\mathcal{R}$ satisfies the 3-fold intersection property.*

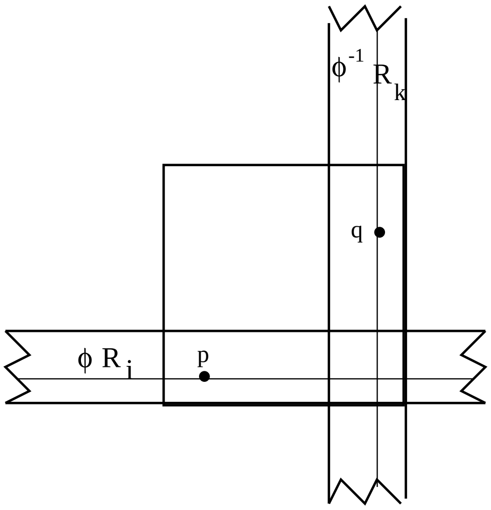


FIGURE 7.IV. 3-fold Intersection Property

PROOF. Let p $\in \phi R_i \cap R_j \neq \emptyset$ and q $\in R_j \cap \phi^{-1}R_k \neq \emptyset$. Then by transversality $v_j(\mathrm{q}) \cap h_j(\mathrm{p}) \neq \emptyset$. Futhermore, $v_j(\mathrm{q}) \cap h_j(\mathrm{p}) \subset R_j$ and $v_j(\mathrm{q}) \cap h_j(\mathrm{p}) \subset \phi^{-1}v_k(\phi\mathrm{q}) \cap \phi h_i(\phi^{-1}\mathrm{p}) \subset \phi^{-1}R_k \cap \phi R_i$: So $\phi R_i \cap R_j \cap \phi^{-1}R_k \neq \emptyset$. Thus we have

$$R_i \cap \phi^{-1}R_j \neq \emptyset, R_j \cap \phi^{-1}R_k \neq \emptyset \Rightarrow R_i \cap \phi^{-1}R_j \cap \phi^{-2}R_k.$$

□

COROLLARY 7.8. *Given a dynamical system (X, ϕ), if a topological partition $\mathcal{R}$ has property M, then $\mathcal{R}^{(n)}$ is satisfies the 3-fold intersection property for $n = 1, 2, \ldots$.*

PROOF. Follows from Corollary 7.6 and Proposition 7.7. $\square$

THEOREM 7.9. *Given a dynamical system (X, ϕ), if a topological partition $\mathcal{R}$ has property M, then $\mathcal{R}$ is Markov.*

PROOF. Follows from Corollaries 7.6, 7.8, and the Bootstrap Lemma 6.3. For instance, $\mathcal{R}^{(n-1)}$ and $\mathcal{R}^{(n)}$ satisfies the 3-fold intersection property. so $\mathcal{R}^{(n-1)}$ satisfies the 4-fold one. Working our way back, we get $\mathcal{R}^{(n-2)}$ satisfies the 5-fold one, etc. Finally, we get that $\mathcal{R}$ satisfies the $(n+2)-$fold intersection property; but this is true for any n. $\square$

We conclude this section with a theorem which is useful in applications to dynamical systems having smooth manifolds as phase spaces. For that theorem boundaries of partition members will play a role. In addition we shall need more topological structure than that provided by mere existence of a pair of transverse partitions.

First we turn our attention to boundaries. In certain problems the burden of establishing the Markov property for a partition via property M can be eased by merely verifying a similar property for boundaries. The reader will get a good illustration of this when we discuss in detail Markov partitions for automorphisms of the two torus.

Employing the usual notation, we have that the boundary of an element R_i in a topological partition $\mathcal{R}$ is given by $\partial R_i \equiv \overline{R_i} - R_i$. We denote the union of all boundaries of elements of $\mathcal{R}$ by $\partial\mathcal{R} \equiv \bigcup_i \partial R_i$. Suppose the boundary ∂R_i of each element of $\mathcal{R}$ is the union of two subsets: one, $\partial_V R_i$ which we shall call the *vertical boundary* of R_i, the other, $\partial_H R_i$, the *horizontal boundary* of R_i. We denote the union of all vertical boundaries of elements of $\mathcal{R}$ by $\partial_V\mathcal{R} \equiv \bigcup_i \partial_V R_i$, and the union of all horizontal ones by $\partial_H\mathcal{R} \equiv \bigcup_i \partial_H R_i$.

DEFINITION 7.10. We say that a topological partition $\mathcal{R}$ has *boundaries satisfying property M* if the following hold for each i :

(1) $\partial R_i = \partial_V R_i \cup \partial_H R_i$,
(2) $\overline{v_i(\mathrm{p})} \cap \partial R_i \subset \partial_H R_i$,
(3) $\overline{h_i(\mathrm{p})} \cap \partial R_i \subset \partial_V R_i$,
(4) $\phi\partial_V\mathcal{R} \subset \partial\mathcal{R}$,
(5) $\phi^{-1}\partial_H\mathcal{R} \subset \partial\mathcal{R}$.

We introduce the additional topological structure needed for the next theorem.

DEFINITON 7.11. We call a metric space R an *abstract rectangle* if it is homeomorphic to the Cartesian product two metric spaces– *i.e.* there exist two metric spaces H, V and a homeomorphism Φ of the Cartesian product $H \times V$ onto R.

Sets with a pair of transverse partitions usually arise in this way. Let $\Phi(\mathrm{p}^1, \mathrm{p}^2) = \mathrm{p}$, where $\mathrm{p} \in R$ and $(\mathrm{p}^1, \mathrm{p}^2) \in H \times V$. Define the following horizontal and vertical sets of R :

$$h(\mathrm{p}) \equiv \Phi\{(x, p^2) : x \in H\},$$

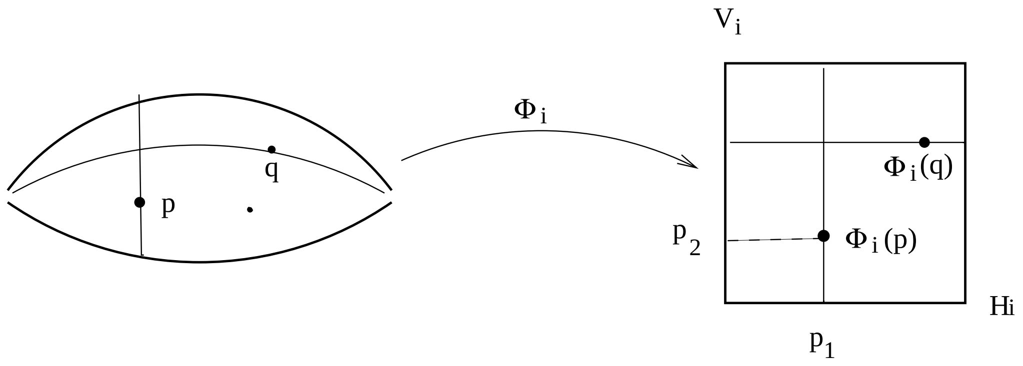


FIGURE 7.V. Abstract rectangle

$$v(\mathrm{p}) \equiv \Phi\{(p^1, y) : y \in V\}.$$

Naturally the two partitions

$$\mathcal{H} = \{h(\mathrm{p}) : \mathrm{p} \in R\},$$

$$\mathcal{V} = \{v(\mathrm{p}) : \mathrm{p} \in R\}$$

of R are transverse since

$$v(\mathrm{p}) \cap h(\mathrm{q}) = \{\Phi(\mathrm{p}^1, \mathrm{q}^2)\} \neq \emptyset.$$

In addition, for each pair of points $\mathrm{p}, \mathrm{q} \in R$ the map $(\mathrm{p}, \mathrm{q}) \mapsto v(\mathrm{p}) \cap h(\mathrm{q})$ is continuous, onto, and maps $h(\mathrm{p}) \times v(\mathrm{q})$ homeomorphically onto R_i. Thus we could have assumed that H, V were subsets of R in the first place. We use the letters we do to designate these subsets in order to suggest horizontal and vertical lines.

THEOREM 7.12. *In a dynamical system (X, ϕ), if each element of a topological partition $\mathcal{R}$ is a connected abstract rectangle, the alignments of which are maintained by ϕ and its inverse respectively, and if $\mathcal{R}$ has boundaries with property M, then $\mathcal{R}$ itself has property M– i.e. $\mathcal{R}$ is a Markov partition.*

PROOF.
We give the proof only for verticals which consists in proving

$$p \in R_i \cap \phi^{-1} R_j \Rightarrow \phi v_i(p) \subset R_j.$$

Our proof involves one proof by contradiction established by means of a second. The main one is a contradiction to the assumption that $\phi v_i(p) \not\subseteq R_j$. The other one contradicts the connectivity of $\phi v_i(p)$, which is a consequence of the the following.

Since R_i homeomorphic to $h_i(p) \times v_i(p)$, the vertical $v_i(p)$ is connected: for otherwise R_i would not be. Therefore, the homeomorphic image $\phi v_i(p)$ is connected as well. Thus, if $\phi v_i(p) \not\subseteq R_j$, then we would have that $\phi v_i(p) \cap \partial R_j \neq \emptyset$. Hence there would exist a point $q \in \phi v_i(p) \cap \partial R_j$ which is also in $\overline{\phi v_i(p) \cap R_j}$: for if not,

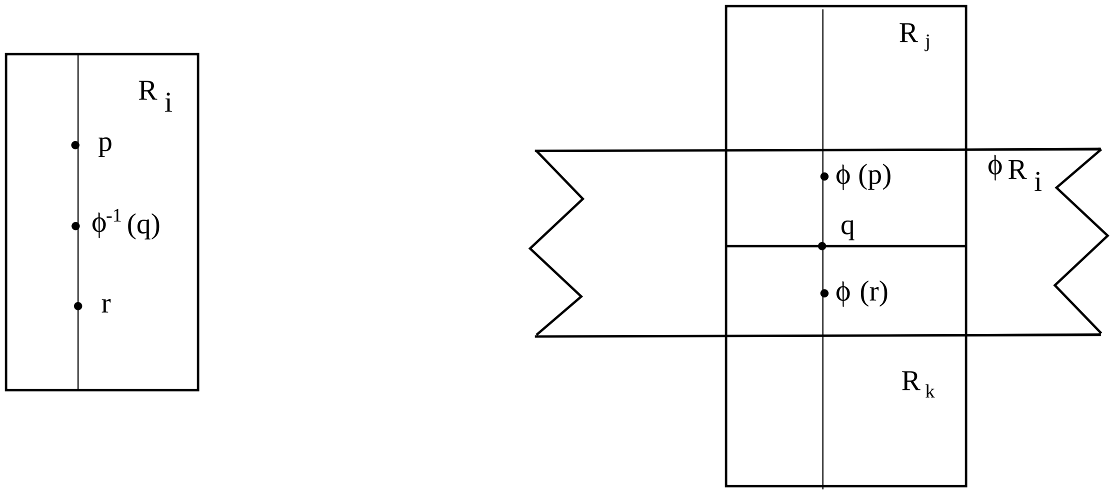


FIGURE 7.VI. Impossible boundary picture

then there would be an open set $U \supset \phi v_i(p) \cap \partial R_j$ such that $U \cap \phi v_i(p) \cap R_j = \emptyset$, and the open sets R_j and $U \cup (X - \overline{R_j})$ would disconnect $\phi v_i(p)$.

By preservation of alignments, we have $\phi v_i(p) \cap R_j \subset v_j(\phi p)$ so that $q \in \overline{v_j(\phi p)}$. Thus $q \in \partial_H \mathcal{R}$ from which follows by property M for boundaries that $\phi^{-1}(q) \in \partial\mathcal{R}$. However, $\phi^{-1}(q) \in v_i(p) \subset R_i$, which contradicts $R_i \cap \partial R_i = \emptyset$.

□

We now turn our attention to 2-dimensional toral automorphisms in generality. While there exists non-measurable automorphisms, for us toral automorphisms will mean continuous ones.

8. Markov Partitions for Automorphisms of the 2-torus.

Let $X = \mathbb{R}^n/\mathbb{Z}^n$ be the n-dimensional torus and A a $n \times n$ matrix with integer entries and determinant ± 1. Such a matrix defines an automorphism of the n-torus in the manner described in Section 2.3. The set of such matrices forms a group called the *general linear group* $GL(n, \mathbb{Z})$. Both a matrix A and the automorphism ϕ it defines are called *hyperbolic*, if A has no eigenvalue of modulus one.

We shall devote the rest of this section to the two dimensional case; *i.e.* $n = 2$. Let

$$\mathrm{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in GL(2, \mathbb{Z}).$$

Eigenvalues of A are are the solutions of the quadratic equation

$$x^2 - (\mathrm{traceA})x + \mathrm{detA} = 0.$$

Here hyperbolicity means that A has two distinct eigenvalues, say λ and μ which are irrational numbers. Since $\lambda\mu = \det A = \pm 1$, we can assume that $|\lambda| > 1$ and

$|\mu| < 1$. An easy calculation shows that the row vectors

$$\begin{aligned} \mathrm{v}_\lambda &= (c, \lambda - a) \\ \mathrm{v}_\mu &= (c, \mu - a) \end{aligned}$$

are eigenvectors associated with λ and with μ respectively. The action of A on a vector v is to contract its v_μ-component by μ and expand its v_λ-component by λ. Directions may or not be reversed depending on the signs of the eigenvalues. We refer to the direction of v_λ as the *expanding* direction and that of v_μ as the *contracting* one. Finally let ℓ_λ be a line through the origin in the expanding direction and ℓ_μ the one in the contracting direction. We call these lines, which are invariant under the action of A on the plane, the *expanding* and *contracting eigen-line* respectively. The slopes of these lines are $m_\lambda = (\lambda - a)/c$ and $m_\mu = (\mu - a)/c$. From these formulae one sees that these lines pass through no lattice points other than the origin: for if they did then the slopes m_λ and m_μ would be rational numbers and so would λ and μ.

THEOREM 8.1. *A hyperbolic toral automorphism is expansive.*

PROOF. Let ϕ be an automorphism and $\mathrm{p}, \mathrm{q} \in X$ be any two different points of the two dimensional torus. Let $c = |\mu|/8$. We shall show that there exists $n \in \mathbb{Z}$ such that $d(\phi^n \mathrm{p}, \phi^n \mathrm{q}) > c$. By translation invariance of the metric we have $d(\mathrm{p}, \mathrm{q}) = d(\{0\}, \mathrm{p} - \mathrm{q})$: so it suffices to show that $d(\{0\}, \phi^n \mathrm{r}) > c$, for any $\mathrm{r} \neq \{0\}$ in X.

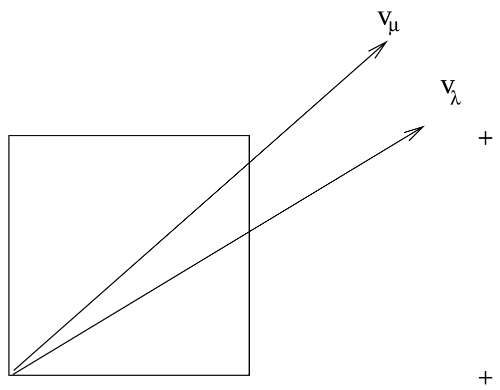


FIGURE 8.I. A fundamental region and eigen-directions of A

We take the torus to be given by the fundamental region

$$X = \{(x, y) : |x| \leq 1/2,\ |y| \leq 1/2\}$$

with the appropriate boundary identifications. In this region the metric on the torus coincides with the Euclidean one: namely, $d(\mathrm{p}, \mathrm{q}) = ||\mathrm{p} - \mathrm{q}||$ where $||(x, y)|| =$

$\sqrt{x^2+y^2}$. Let $\mathrm{r} = \mathrm{p} - \mathrm{q} \neq (0,0)$. Let r_λ and r_μ be the $\mathrm{v}_\lambda-$component and the $\mathrm{v}_\mu-$component of r respectively. Then from the triangle inequality

$$|\lambda^n \mathrm{r}_\lambda| - |\mu^n \mathrm{r}_\mu| \leq ||\phi^n \mathrm{r}|| \leq |\lambda^n \mathrm{r}_\lambda| + |\mu^n \mathrm{r}_\mu|.$$

One of the components $\mathrm{r}_\lambda, \mathrm{r}_\mu$ is not zero. We can assume that $\mathrm{r}_\lambda \neq 0$: otherwise replace ϕ by ϕ^{-1} in the argument. We can also assume that $|\mu| < \frac{1}{2}$: for, if not, replace ϕ by ϕ^k for large enough k. If $||\mathrm{r}|| > \frac{|\mu|}{4}$, then $||\phi^n \mathrm{r}|| > c$ for $n = 0$. If $||\mathrm{r}|| \leq \frac{|\mu|}{4}$, choose $n \geq 1$ such that

$$\frac{|\mu|^{n+1}}{4} \leq |r_\lambda| \leq \frac{|\mu|^n}{4}.$$

Then the following inequalities inequality show that $\phi^n \mathrm{r}$ is in the fundamental region and that $||\phi^n \mathrm{r}|| > c$.

$$|\lambda^n \mathrm{r}_\lambda| + |\mu^n \mathrm{r}_\mu| \leq \frac{1}{4} + \frac{|\mu|^{n+1}}{4} < \frac{1}{2},$$

$$\frac{|\mu|}{8} \leq \frac{|\mu|}{4} - \frac{|\mu|^{n+1}}{4} \leq |\lambda^n \mathrm{r}_\lambda| - |\mu^n \mathrm{r}_\mu|.$$

□

Remark. We now have all the ingredients for a formal proof that the toral automorphism of example 3.3 enjoys the conclusions of theorem 6.5 about representing dynamical systems by topological Markov shifts. In this case the automorphisms ϕ be given by the matrix

$$\mathrm{A} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

and the topological Markov shift (Σ_G, σ) by the edge-graph G of figure 3.VIII. Theorem 8.1 shows that ϕ is expansive, a necessary item in the hypothesis of the main theorem, Theorem 6.5. Theorem 7.13 can be applied to show that the partition in example 3.3 is Markov, another necessary item. Finally in 3.3 we have already shown that the partition in question is a generator, the remaining requirement of the main theorem.

Returning our attentions to the general case, we shall exploit one of the properties of a hyperbolic automorphisms: namely, a matrix in the group $GL(2,\mathbb{Z})$ specifying a hyperbolic automorphism is conjugate to another one all of whose entries bear the same sign. More specifically we have the following theorem of Williams [W] proved by entirely elementary methods of plane geometry.

Theorem 8.2. *Let* $\mathrm{A} \in GL(2,\mathbb{Z})$ *be hyperbolic. Then there exists* $\mathrm{C}, \mathrm{P} \in GL(2,\mathbb{Z})$ *such that* $\mathrm{CAC}^{-1} = \epsilon \mathrm{P}$ *where* $\epsilon = \pm 1$, *the choice of sign being the same as that of* λ, *and the entries of the matrix*

$$\mathrm{P} = \begin{pmatrix} p & q \\ r & s \end{pmatrix}$$

are non-negative.

Proof.

Choose a pair of lattice points (α, β) , (γ, δ) such that

(i) The angle $(\alpha, \beta)(0,0)(\gamma, \delta)$ is acute and ℓ_λ lies in it but ℓ_μ does not.

(ii) The closed parallelogram with vertices the origin, the two lattice points, and the lattice point $(\alpha+\gamma, \beta+\delta)$ contains no other lattice point.

This can always be done. One way is to use continued fractions to approximate slopes of lines, a discussion of which shall be deferred to a remark. An even more elementary way is the following.

First, chose initial lattice points (α, β) and (γ, δ) so close to ℓ_λ that (i) is satisfied and no other lattice points lie between them and the origin on a direct line. Then if the parallelogram in (ii) contains another lattice point in its interior, connect it to the origin with a line segment. Form a new pair of lattice points by taking the closest lattice point to the origin on this segment and selecting the one from the previous pair for which (i) holds. Continue this process until condition (ii) is satisfied.

Condition (ii) is equivalent to the area of the parallelogram equalling 1.

Consider the linear map given by the matrix

$$\mathrm{C} = \begin{pmatrix} \gamma & \delta \\ \alpha & \beta \end{pmatrix}.$$

Since

$$(0,1)\mathrm{C} = (\alpha, \beta)$$

$$(1,0)\mathrm{C} = (\gamma, \delta),$$

C maps the principal fundamental region–namely, the closed unit square–onto the parallelogram. We shall show that C provides the sought after conjugating transformation.

The two lines $\overline{\ell}_\lambda, \overline{\ell}_\mu$ that are the images under C^{-1} of ℓ_λ, ℓ_μ are the expanding and contracting eigen-lines for each of the transformations $\pm\mathrm{CAC}^{-1}$. Choose P to be the one that preserves the orientation of $\overline{\ell}_\lambda$, the choice of sign being that of λ. This means that the matrix P has $|\lambda|$ as the expanding eigenvalue and either $\pm\mu$ as the contracting. Furthermore, because C^{-1} maps the parallelogram onto the unit square and ℓ_λ passes through the parallelogram while ℓ_μ does not, the line $\overline{\ell}_\lambda$ passes through the first quadrant and the line $\overline{\ell}_\mu$ the second.

To prove that P is non-negative, or equivalently, that P maps the first quadrant into itself, we must just show $(0,1)\mathrm{P}$ and $(1,0)\mathrm{P}$ lie in the first quadrant. We shall give the proof only for $(0,1)\mathrm{P}$: the arguments which follow work equally well for the other lattice point $(1,0)\mathrm{P}$.

There are two cases depending whether the contracting eigenvalue μ of P is positive or negative. In the second case ($\mu < 0$), the linear map P reflects the first quadrant about the eigen-line $\overline{\ell}_\lambda$, while in the first case ($\mu > 0$) no reflection takes place. Let c denote the origin, a the lattice point $(0,1)$, and a$'$ its image under P.

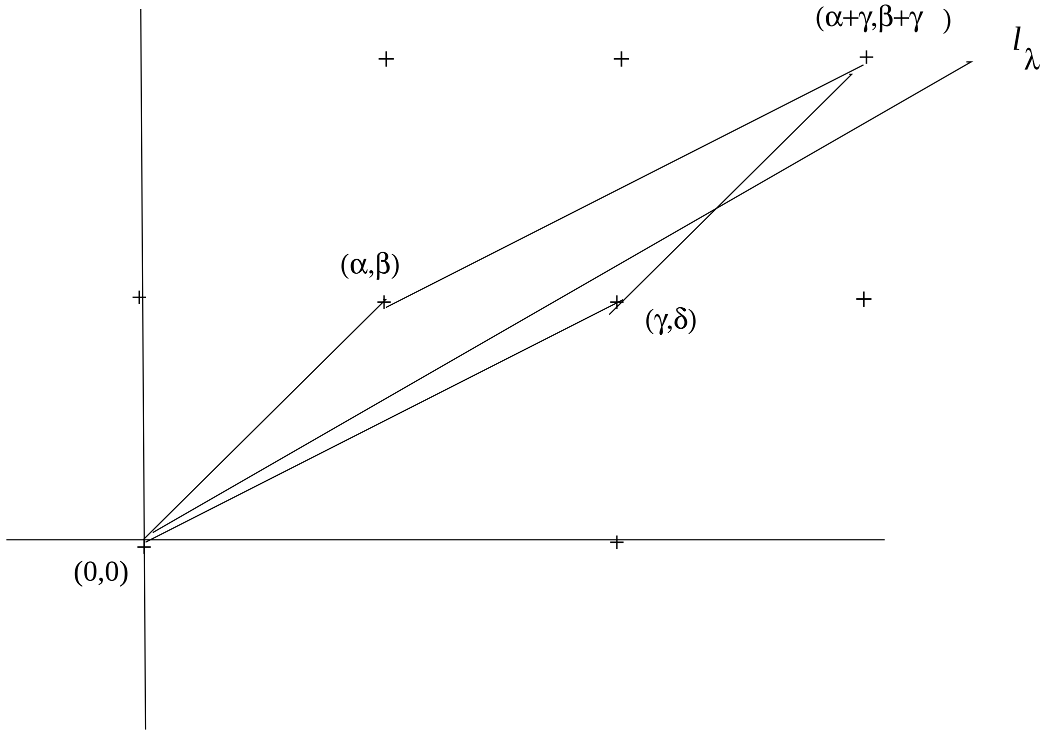


FIGURE 8.II. Parallelogram and expanding direction

Let b be the projection of a on the line $\overline{\ell}_\lambda$ in the direction parallel to $\overline{\ell}_\mu$ and b$'$ its image under P. The point b$'$ is also the projection of a$'$ on the line $\overline{\ell}_\lambda$.

We deal first with the case without reflection. Let $\overline{\ell}'_\lambda, \overline{\ell}'_\mu$ be the lines through a parallel to $\overline{\ell}_\lambda, \overline{\ell}_\mu$ respectively. The notation $|\mathrm{pq}|$ stands for the length of the line segment with end-points p, q. On one hand, since $|a'b'| = |\mu|^{-1} \cdot |\mathrm{ab}| < |\mathrm{ab}|$, the point a$'$ lies between the lines $\overline{\ell}'_\lambda, \overline{\ell}_\lambda$. On the other, since $|\mathrm{cb}'| = |\lambda| \cdot |\mathrm{cb}| > |\mathrm{cb}|$, the point a' lies to the left of $\overline{\ell}'_\mu$. The region bounded by these three lines, in which a' thus lies, is contained in the first quadrant.

For the case with reflection, let c' be the intersection of the line $\overline{\ell}_\lambda$ and the vertical through a$'$. Suppose that a$'$ belongs to the fourth quadrant but not the first. The point a' being a lattice point implies that $|\mathrm{a'c'}| > 1$. However, because triangle c$'$a$'$b$'$ is similar to cab, $|\mathrm{a'c'}| = |\mu| \cdot |\mathrm{ac}| < |\mathrm{ac}| < 1$, a contradiction.
□

REMARK. By means of continued fractions we can somewhat augment the conclusion of Theorem 8.2: namely, *we can conjugate so that the following two conditions to hold simultaneously.*

(1) $\overline{m}_\mu < -1$,

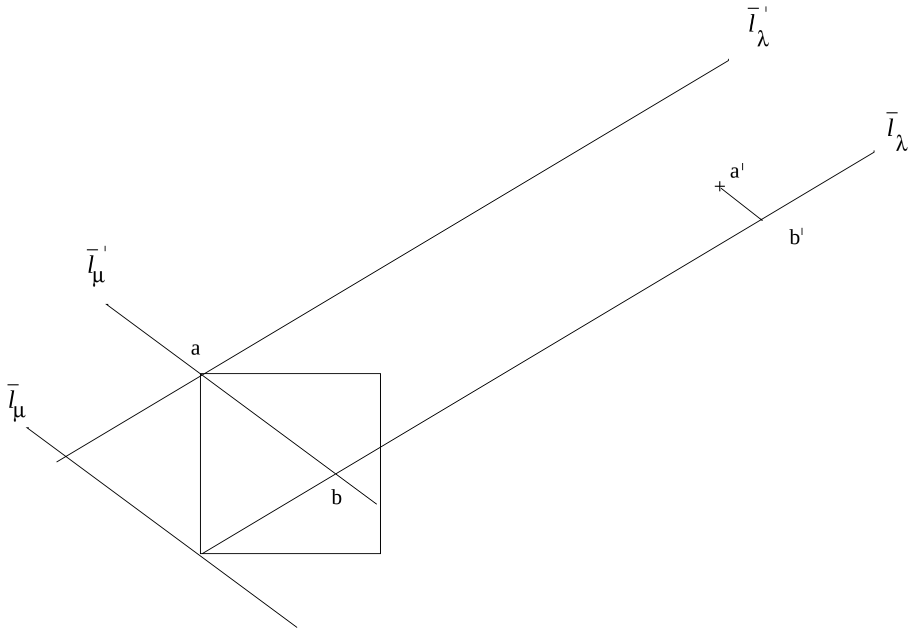


FIGURE 8.III. Geometrical figure for proof without reflection

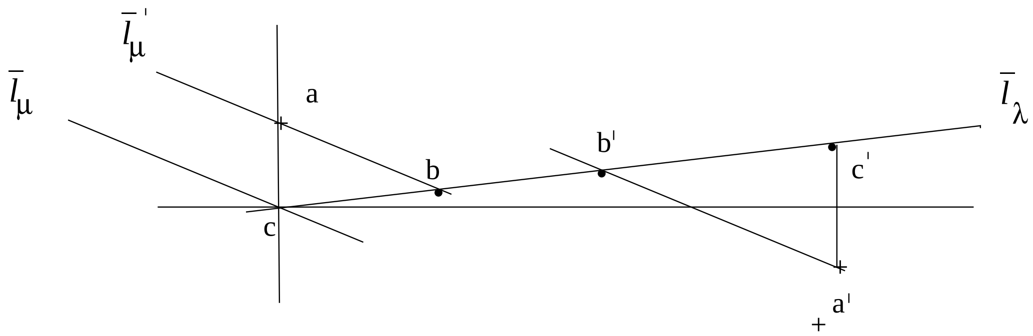


FIGURE 8.IV. Geometrical figure for proof with reflection

(2) $0 < \overline{m}_\lambda < 1$,

where $\overline{m}_\mu$ *is the slope of the contracting eigenline* $\overline{\ell}_\mu$ *and* $\overline{m}_\lambda$ *is the slope of the expanding one* $\overline{\ell}_\lambda$ *for* P.

The first inequality indicates that the contracting eigen-line for P passes through the second quadrant between the lattice points $(0,1)$ and $(-1,1)$; and the second that the expanding one passes through the first quadrant under the lattice point $(1,1)$. For Theorem 8.4, the main one of this section, one does not need more than what is provided by Theorem 8.2. These extra properties make life a little less difficult. The first one makes Figure 8.VI easier to draw. The second one obviates repeating proofs covering slightly different geometrical figures. Not taking advantage of it multiplies the number of cases in the proof, and we shall have enough of them as it is. Since we shall not be using the full strength of this remark one can skip the remainder of it and proceed to directly Theorem 8.4. We shall be using the second property which is very easy to achieve by itself.

From the theory of continued fractions, we know that every irrational number can be written uniquely as an infinite continued fraction $[a_0, a_1, \dots] = a_0 + 1/(a_1 + 1/\dots)$ where $a_n \in \mathbb{Z}$ for all n and $a_n > 0$ for $n > 0$. In addition, the continued fraction of a quadratic surd has a periodic tail: namely the tail can be written as $[\overline{b_1, \dots, b_m}] > 1$, where the overbar means infinite repetition of b_1 through b_m. In [ATW] the following was proved.

THEOREM 8.3. *Let* $\mathrm{A} \in GL(2, \mathbb{Z})$ *be hyperbolic. The slope* m_λ*, being a quadratic surd, can be written* $m_\lambda = [a_0, a_1, \dots, a_n, \overline{b_1, \dots, b_m}]$*, where* m *is as small as possible. If*

$$\mathrm{C} = \begin{pmatrix} 0 & 1 \\ 1 & a_0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & a_1 \end{pmatrix} \cdots \begin{pmatrix} 0 & 1 \\ 1 & a_n \end{pmatrix},$$

then $\mathrm{CAC}^{-1} = \epsilon \mathrm{P}$ *where*

$$\mathrm{P} = \left(\begin{pmatrix} 0 & 1 \\ 1 & b_1 \end{pmatrix} \cdots \begin{pmatrix} 0 & 1 \\ 1 & b_m \end{pmatrix} \right)^N,$$

for some positive integer N *and* ϵ *is the same as in Theorem 8.2. Furthermore, the slopes the eigen-lines of* P *satisfy* $\overline{m}_\lambda = [\overline{b_1, \dots, b_m}] > 1$ *and* $-1/\overline{m}_\mu = [\overline{b_m, \dots, b_1}] > 1$.

The matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \mathrm{P} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

achieves the above two conditions in the remark.

THEOREM 8.4. *Let* ϕ *be a toral automorphism whose defining matrix is either* P *or* $-\mathrm{P}$ *where*

$$\mathrm{P} = \begin{pmatrix} p & q \\ r & s \end{pmatrix}$$

is a hyperbolic matrix in $GL(2, \mathbb{Z})$*, with non-negative entries. Then there exists a Markov generator* $\mathcal{R}^*$ *for* ϕ*, the members of which are of parallelograms. The associated Markov shift is given by a directed graph also specified by* P *: i.e., the edge graph with connections given by* P *consists of two vertices labelled I and II with* p *directed edges from I to itself,* q *from I to II,* r *from II to I, and* s *from II to itself.*

PROOF. We shall assume that the expanding eigen-line of P as given by Theorem 8.2 passes under the point $(1, 1)$; if not,conjugate P by the matrix

$$\mathrm{E} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

which reflects the first quadrant about the line $y = x$.

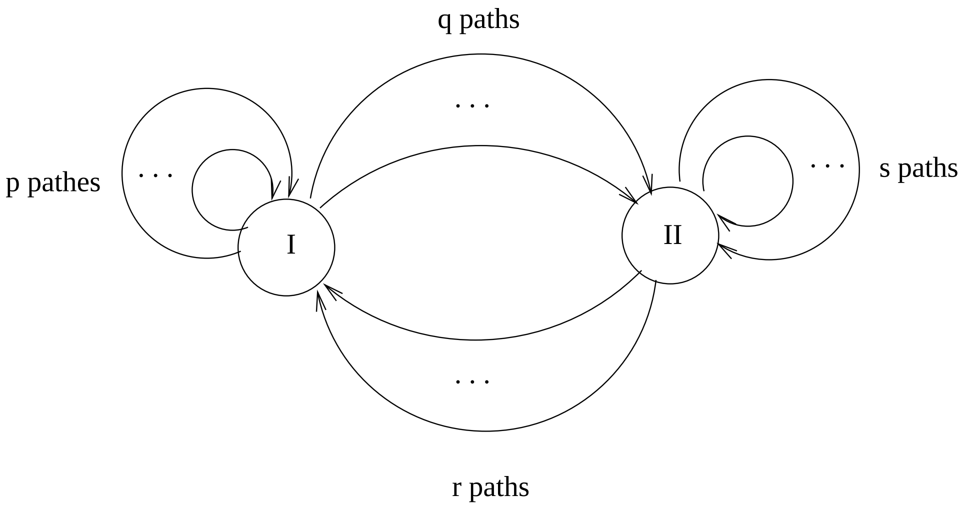


FIGURE 8.V. Edge graph defined by P

Before proceeding in earnest, we need some notation. Dropping the bars, we now let v_λ, v_μ be the expanding and contracting eigen-vectors of P and ℓ_λ, ℓ_μ the corresponding eigenlines through the origin. We denote lines parallel to these through a point p by assigning p as a superscript. For example, $\ell_\lambda^{(0,1)}$ denotes the line through $(0,1)$ parallel to ℓ_λ, etc.

We define the following points as depicted in Figure 8.VI:

$$
\begin{aligned}
\mathrm{o} &\equiv (0,0)\\
\mathrm{o}' &\equiv (1,0)\\
\mathrm{o}'' &\equiv (1,1)\\
\mathrm{o}''' &\equiv (0,1)\\
\mathrm{a} &\equiv \ell_\mu \cap \ell_\lambda\\
\mathrm{a}' &\equiv \ell_\mu^{(1,0)} \cap \ell_\lambda^{(1,0)}\\
\mathrm{a}'' &\equiv \ell_\mu^{(1,1)} \cap \ell_\lambda^{(1,1)}\\
\mathrm{a}''' &\equiv \ell_\mu^{(0,1)} \cap \ell_\lambda^{(0,1)}\\
\mathrm{b} &\equiv \ell_\mu \cap \ell_\lambda^{(-1,0)}\\
\mathrm{b}' &\equiv \ell\mu^{(1,0)} \cap \ell_\lambda = b + (1,0)\\
\mathrm{b}'' &\equiv \ell_\mu^{(1,1)} \cap \ell_\lambda^{(0,1)} = \mathrm{b} + (1,1)
\end{aligned}
\tag{8.5}
$$

$$\begin{aligned}
\mathrm{c} &\equiv \ell_\mu \cap \ell_\lambda^{(0,1)}\\
\mathrm{c}' &\equiv \ell_\mu^{(1,0)} \cap \ell_\lambda^{(1,1)}\\
\overline{\mathrm{c}} &\equiv \ell_\mu^{(0,-1)} \cap \ell_\lambda\\
\overline{\mathrm{d}} &\equiv \ell_\mu \cap \ell_\lambda^{(-1,1)}\\
\mathrm{d}' &\equiv \ell_\mu^{(1,0)} \cap \ell_\lambda^{(0,1)} = \overline{\mathrm{d}} + (1,0)\\
\mathrm{d}^* &\equiv \mathrm{d}' - (0,1).
\end{aligned}$$

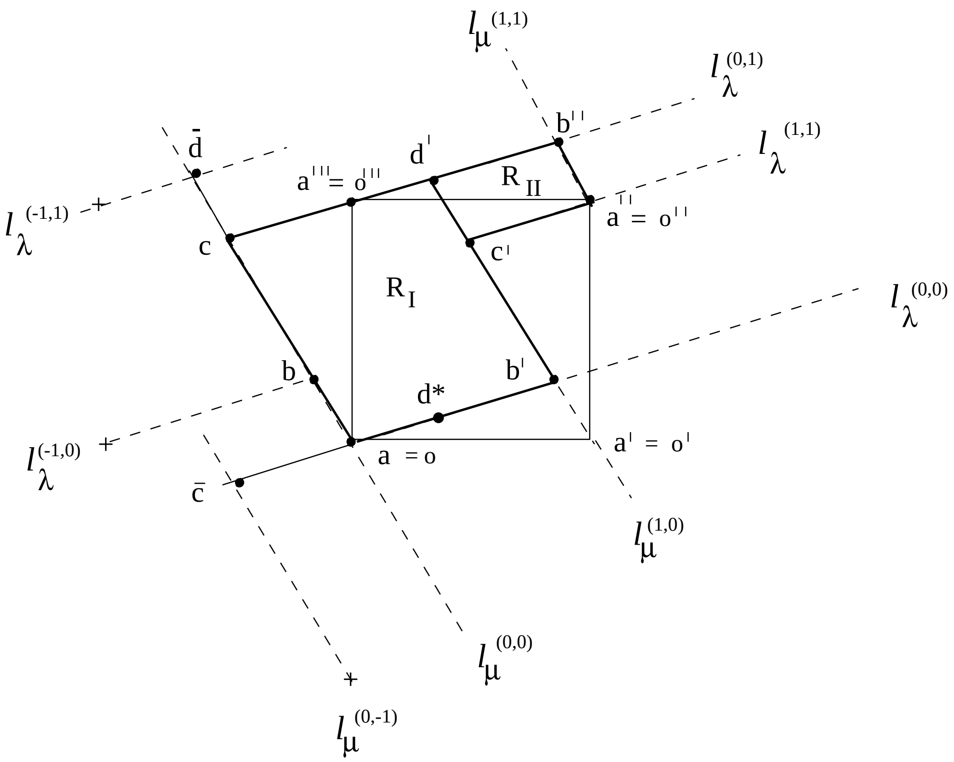


FIGURE 8.VI. A remarkable fundamental region

We have drawn Figure 8.VI as if the first statement in the remark following Theorem 8.2 holds. This places the point c′ in the unit square. Since we are not using this condition, c′ could appear anywhere to the left of the line $x = 1$ in the strip between the lines $y = 0$ and $y = 1$.

Let R_I be the interior of parallelogram acd′b′ and R_{II} the interior of parallelogram c′d′b″a″.

The closed set $\overline{R_I} \cup \overline{R_{II}}$, as we shall show, is a fundamental region, which we shall call the *principal Markov* one. As drawn in Figure 8.VI, this set is equivalent modulo $\mathbb{Z}^2$ to the unit square by sliding $\triangle$aa‴c one unit to the right and $\triangle$d′b″a″ one unit down. But in general we need a slightly more elaborate proof.

First, no two points in the interior of $\overline{R_I} \cup \overline{R_{II}}$ are equivalent because $R_I \cup R_{II}$ is disjoint from its four neighboring unit translates which totally bound it. Second, the set of all $\mathbb{Z}^2-$translates of $\overline{R_I} \cup \overline{R_{II}}$ covers the plane because all integral horizontal translates of the union of this set with its unit downward vertical translate covers the infinite strip between the lines $y = 0$ and $y = 1$ and all integral vertical translates of the strip covers the plane.

Thus we can view the torus to be the set $X = \overline{R_I} \cup \overline{R_{II}}$ with points on the boundary identified by lattice translations.

The proof of the theorem involves four cases: two subcases arise for each matrix $\pm$P depending whether μ is positive or negative. These amount to the four possible combinations,

$$(\lambda, \mu) = (\pm|\lambda|, \pm|\mu|),$$

of signs for the eigenvalues of the matrix representing ϕ. The simplest case is when both λ and μ are positive. Things are more difficult when either is negative, especially μ. So that the proof appears less tedious, we divide that for each case into five steps.

Case I: $\lambda > 0, \mu > 0$.

STEP 1. *The family $\mathcal{R} = \{R_I, R_{II}\}$ is a Markov partition.*

Our aim is to show that $\mathcal{R}$ satisfies the hypothesis of Theorem 7.12.

By virtue of their construction as parallelograms the members R_I, R_{II} of $\mathcal{R}$ are connected. These open parallelograms are obviously abstract rectangles, each being homeomorphic to the Cartesian product of two of open intervals. The two in question are a pair intersecting sides of a parallelogram minus endpoints. Horizontals and verticals of $R_i, i = I, II$, are given by

$$v_i(\mathrm{p}) \equiv \ell_\mu^{\mathrm{p}} \cap R_i,$$

$$h_i(\mathrm{p}) \equiv \ell_\lambda^{\mathrm{p}} \cap R_i.$$

Since P defines a linear transformation of the plane, the image of a line parallel to an eigen-line is another line parallel to the same eigen-line: so alignment of verticals and horizontals is maintained by ϕ and ϕ^{-1} : namely,

$$\mathrm{p} \in R_i \cap \phi^{-1} R_j \Rightarrow R_j \cap \phi v_i(\mathrm{p}) \subset v_j(\phi p); \tag{1}$$

$$\mathrm{p} \in R_i \cap \phi R_j \quad \Rightarrow R_j \cap \phi^{-1} h_i(\mathrm{p}) \subset h_j(\phi^{-1} p) \tag{2}$$

for $i, j = I, II$.

Next we verify that the boundaries of these of members satisfy property M of Definition 7.10 which entails five items. The first of these concerns dividing ∂R_I and ∂R_{II} into vertical and horizontal pieces. As shown in Figure 8.VI we have

$$\partial R_I = \partial_V R_I \cup \partial_H R_I,$$

$$\partial R_{II} = \partial_V R_{II} \cup \partial_H R_{II},$$

where

$$\begin{aligned}\partial_V R_I &\equiv \mathrm{ac} \cup \mathrm{b'd'},\\ \partial_H R_I &\equiv \mathrm{ab'} \cup \mathrm{cd'},\\ \partial_V R_{II} &\equiv \mathrm{c'd'} \cup \mathrm{a''b''},\\ \partial_H R_{II} &\equiv \mathrm{c'a''} \cup \mathrm{d'b''},\end{aligned}$$

each being the union of two line segments. Utilizing the boundary identifications, we have that $\partial\mathcal{R}$ consists of two transverse line segments intersecting at the origin: namely,

$$\begin{aligned}\partial_H \mathcal{R} &= \overline{\mathrm{c}}\mathrm{b'} \subset \ell_\lambda^a = \ell_\lambda,\\ \partial_V \mathcal{R} &= \mathrm{a}\overline{\mathrm{d}} \subset \ell_\mu.\end{aligned} \tag{8.6}$$

It is clear from Figure 8.VI that (2) and (3) of 7.10 are satisfied: namely,

$$\begin{aligned}\overline{v_i(\mathrm{p})} \cap \partial R_i &\subset \partial_H R_i,\\ \overline{h_i(\mathrm{p})} \cap \partial R_i &\subset \partial_V R_i,\end{aligned}$$

From (8.6) and the property that the restrictions of P and P^{-1} respectively to the lines ℓ_λ and ℓ_μ are contractions we get

$$\begin{aligned}\overline{\mathrm{c}}\mathrm{b'} &\subset (\overline{\mathrm{c}}\mathrm{b'})P\\ (\mathrm{a}\overline{\mathrm{d}})P &\subset \mathrm{a}\overline{\mathrm{d}}.\end{aligned} \tag{8.7}$$

We can restate (8.7) as items (4) and (5) of 7.10: namely,

$$\begin{aligned}\phi\partial_V\mathcal{R} &\subset \partial\mathcal{R}_V \subset \partial\mathcal{R},\\ \phi^{-1}\partial_H\mathcal{R} &\subset \partial_H\mathcal{R} \subset \partial\mathcal{R}.\end{aligned}$$

Thus we have established that $\mathcal{R}$ satisfies the hypothesis of theorem 7.12: so it is Markov partition.

Remark. At this point it may be instructive to remark on our definition of topological partition. This example illustrates the advantage of using open sets over their closures as members of such a partition. For one thing, $R_I \neq \overline{R_I}^o$ and $R_{II} \neq \overline{R_{II}}^o$. For another, while R_I and R_{II} are abstract rectangles, their closures are not. For instance, R_I is homeomorphic to a Cartesian product: namely, the product of the line segment ab′ minus the end-points with the segment *ac* minus the end-points. However, $\overline{R_I}$ is not homeomorphic to the Cartesian product of ab′ with ac, since part of the segment ab′ is in the boundary and part isn't. Furthermore, $\partial_V\mathcal{R}$ and $\partial_H\mathcal{R}$ are connected line segments which get mapped into themselves under ϕ and ϕ^{-1} respectively. This would not be the case if $\mathcal{R} = \{\overline{R_I}, \overline{R_{II}}\}$. If one's definition of topological partition involves closures of open sets as members rather than open sets, then one is forced into somewhat greater contortions in order to achieve the same results.

The Markov partition $\mathcal{R}$ is not necessarily a generator. The trouble is that, while the members of $\mathcal{R}$ are connected, those of $\mathcal{R}^{(2)} = \mathcal{R} \vee \phi^{-1}\mathcal{R}$ may not be in which case nonempty sets of the form $\bigcap_{n=0}^{\infty} \overline{\bigcap_{-n}^{n} \phi^{-k} R_{s_k}}$ may consist of more than one point. To overcome this let us examine the sets $R_i \cap \phi^{-1}R_j \in \mathcal{R} \vee \phi^{-1}\mathcal{R}$. As we shall see each consists of a union of disjoint open parallelograms, the number of which is given by the matrix P. A remedy is immediately suggested.

STEP 2. *The family* $\mathcal{R}^* = \{R_k^* : 1 \le k \le N^*\}$ *consisting of all connected components of the sets* $R_i \cap \phi^{-1} R_j \in \mathcal{R} \vee \phi^{-1}\mathcal{R}$ *is Markov partition.*

Once again we must show that the members of $\mathcal{R}^*$ satisfy the five items of Definition 7.10.

In the universal cover the image (R_i)P is an open parallelogram that has been stretched by a factor of λ in the v_λ-direction and shrunk by a factor μ in the v_μ-direction. This parallelogram passes through various Markov fundamental regions; and in each one it when it intersects a parallelogram equivalent to R_j, the intersection is a parallelogram. No two of these intersections share an equivalent point: for otherwise a violation of (2.1.1(3)) with respect to the fundamental region $(\overline{R_I} \cup \overline{R_{II}})$P would be committed. While the set $\phi h_i(\mathrm{p}) \cap R_j$ may consist of several horizontals, the set $\phi h_i(\mathrm{p}) \cap R_k^*$ is either empty or a single one. Furthermore, if $\phi h_i(\mathrm{p}) \cap R_k^* \ne \emptyset$ for one horizontal $h_i(\mathrm{p})$ of R_i, then $\phi h_i(\mathrm{p}') \cap R_k^* \ne \emptyset$ for any other horizontal $h_i(\mathrm{p}')$ in R_i. Back on the torus the various non-equivalent parallelograms represent disjoint connected sets which we have labelled $R_k^*, 1 \le k \le N^*$. Being parallelograms, these sets are abstract rectangles.

Alignment of verticals and horizontals for these members of $\mathcal{R}^*$ is maintained by ϕ and ϕ^{-1} for the same reason it is for members of $\mathcal{R}$.

Regarding boundaries we have that

$$\begin{aligned} \partial_H \mathcal{R}^{(2)} &= \partial_H \mathcal{R}^* = (\mathrm{c}''\mathrm{b}')\mathrm{P}, \\ \partial_V \mathcal{R}^{(2)} &= \partial_V \mathcal{R}^* = \mathrm{ad}. \end{aligned} \tag{8.8}$$

Thus, just as for $\mathcal{R}$, we have

$$\begin{aligned} \phi \partial_V \mathcal{R}^* &= (\mathrm{ad})P \subset ad = \partial \mathcal{R}_V^* \subset \partial \mathcal{R}^*, \\ \phi^{-1} \partial_H \mathcal{R}^* &= \mathrm{c}''\mathrm{b}' \subset (\mathrm{c}''\mathrm{b}')\mathrm{P} = \partial_H \mathcal{R}^* \subset \partial \mathcal{R}^*. \end{aligned}$$

The partition $\mathcal{R}^*$ satisfies the hypothesis of Theorem 7.12: so it is Markov. By reasoning as before, a parallelogram of the form (R_k^*)P passes through various fundamental regions; and in each one it when it intersects a parallelogram equivalent to R_l^*, the intersection is a parallelogram, no two of which intersections share an equivalent point. Hence, a nonempty set of the form $\phi R_i^* \cap R_j^*$ is a single connected parallelogram.

If R_k^* is one of the parallelograms in $\phi R_i \cap R_j$, then the length of its v_λ-dimension is the same as that of R_j while the length of its v_μ-dimemsion is $|\mu|$ times that of R_i. Similarly, a non-empty set of the form $\phi R_i^* \cap R_j^*$ is a single connected open parallelogram, the length of its v_λ-dimension being the same as that of R_j^* and the length of its v_μ-dimension μ times that of R_i^*.

Likewise, a non-empty set of the form $\bigcap_{-n}^{n} \phi^{-k} R_{s_k}^*$ is a single connected open parallelogram, the length of its v_λ-dimension is $|\mu|^n$ times as that of $R_{s_{-n}}^*$ and the length of its v_μ-dimension $|\mu|^n$ times that of $R_{s_n}^*$. From this we have

$$d\left(\bigvee_{-n}^{n} \phi^k \mathcal{R}^*\right) = d(\mathcal{R})/|\mu|^{n+1} \to 0, \text{as } n \to \infty.$$

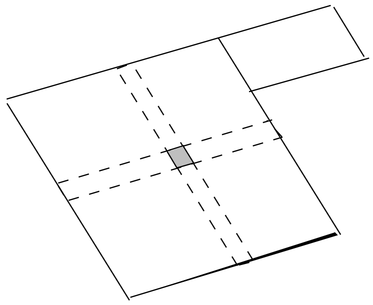


FIGURE 8.VII. A nonempty parallelogram of the form $\bigcap_{-n}^{n} \phi^{-k} R^*_{s_k}$

In other words, $\mathcal{R}^*$ is a generator.

STEP 4. *Let $\#(\phi R_i \cap R_j)$ denote the number of disjoint parallelograms in the intersection $\phi R_i \cap R_j$.*

$$
\begin{aligned}
\#(\phi R_i \cap R_I) &\equiv \text{ number of lines } x = 0, x = 1, \ldots, \text{ traversed by } (R_i)P; \\
\#(\phi R_i \cap R_{II}) &\equiv \text{ number of lines } y = 1, y = 2, \ldots, \text{ traversed by } (R_i)P,
\end{aligned}
\tag{8.9}
$$

which leads to

$$
\begin{aligned}
\#(\phi R_I \cap R_I) &= x\text{-coordinate of } (0,1)\mathrm{P} &&= p; \\
\#(\phi R_{II} \cap R_I) &= x\text{-coordinate of } (1,1)\mathrm{P} - (1,0)\mathrm{P} &&= q; \\
\#(\phi R_I \cap R_{II}) &= y\text{-coordinate of } (0,1)\mathrm{P} &&= r; \\
\#(\phi R_{II} \cap R_{II}) &= y\text{-coordinate of } (1,1)\mathrm{P} - (1,0)\mathrm{P} &&= s.
\end{aligned}
\tag{8.10}
$$

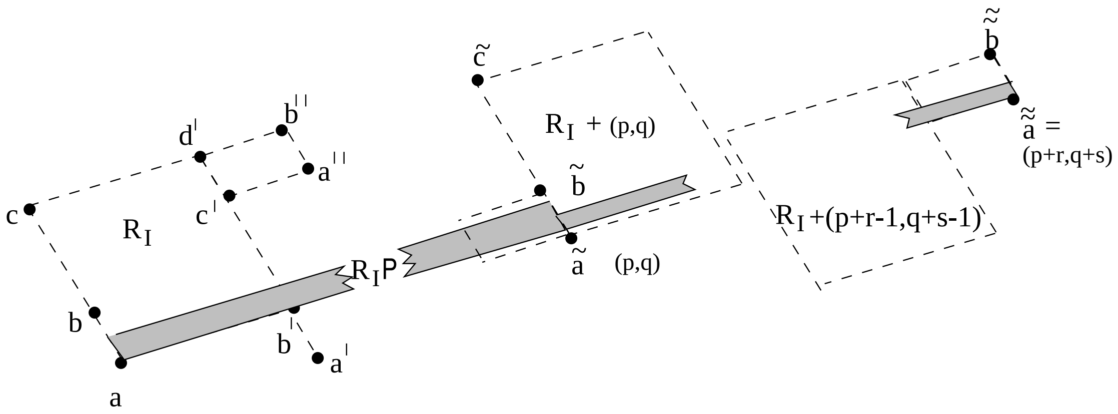


FIGURE 8.VIII. How (R_I)P and (R_{II})P intersect various fundamental regions

To establish Step 4 it is important to understand how $(R_I)P$ and $(R_{II}P$ intersect various Markov fundamental regions particularly how they begin and end. We must show that Figure 8.VIII truly represents the situation: namely the segment $(b'd')$P lies within the segment $\tilde{a}\tilde{b}$. First, if a parallelogram (R_i)P intersects a lattice translate of R_j then it stripes all the way across. As it does do, it passes strictly through the lattice translate without ever straddling any part of a horizontal boundary: for otherwise there would be a violation of the property that $(\overline{R_I} \cup \overline{R_{II}})$P is a fundamental region because $\partial_H(\mathcal{R}\mathrm{P}) \supset \partial_H\mathcal{R}$.

Since P is a contraction on the line ℓ_μ, we have $|(\mathrm{ab})\mathrm{P}| < |\mathrm{ab}|$; and since the point $\mathrm{a} = (0,0)$ is fixed, the point (b)P lies inside the segment ab. It then follows that the point $(\mathrm{b}')\mathrm{P} = (\mathrm{b})\mathrm{P} + (p,q)$ lies inside a lattice translate of ab–that is, it lies on the line $\ell_\mu^{(p,q)}$ strictly between $\tilde{\mathrm{a}} \equiv (p,q) = (\mathrm{a}')\mathrm{P}$ and $\tilde{\mathrm{b}} \equiv (p,q) + \mathrm{b}$. The same is true for $(\mathrm{d}')\mathrm{P}$: for otherwise (R_I)P would straddle part of its own boundary on the line $\ell^{(p-1,q)}$. Thus

$$(\mathrm{ac})\mathrm{P} \subset (\mathrm{ad})\mathrm{P} \subset \mathrm{ab}.$$

and

$$(\mathrm{b}'\mathrm{d}')\mathrm{P} \subset \tilde{\mathrm{a}}\tilde{\mathrm{b}}.$$

So the parallelogram (R_I)P begins (as shown in the leftmost figure of Figure 8.VIII) with its left vertical boundary contained in the segment ab on the line ℓ_μ and ends (as shown in the middle figure of Figure 8.VIII) with its right vertical boundary contained in the segment $\tilde{\mathrm{a}}\tilde{\mathrm{b}}$ on the line $\ell_\mu^{(p,q)}$. We see that the (R_I)P begins by striping across R_I. Its lower horizontal boundary lies on the line ℓ_λ and covers the lower horizontal boundary of R_I. Not shown in the figure is manner by which (R_I)P stripes across the top of fundamental region consisting of $R_I+(r,s-1)$ and $R_{II}+(r,s-1)$ with its upper horizontal boundary contained in the line $\ell_\lambda^{(\mathrm{a}''')\mathrm{P}} = \ell_\lambda^{(r,s)}$. The parallelogram (R_I)P ends by striping through $R_{II} + (p-1, q-1)$.

The other parallelogram (R_{II})P begins where (R_I)P leaves off. Its left vertical boundary is contained in $\tilde{\mathrm{a}}\tilde{\mathrm{b}}$, and its right one in $\tilde{\tilde{\mathrm{a}}}\tilde{\tilde{\mathrm{b}}}$ where $\tilde{\tilde{\mathrm{a}}} \equiv (p+r, q+s)$ and $\tilde{\tilde{\mathrm{b}}} \equiv \mathrm{b} + (p,q)$ as shown in the middle figure of Figure 8.VIII. The first set that (R_{II})P stripes through is $R_I + (p,q)$ and the last $R_{II} + (p+r-1, q+s-1)$. It stripes through $R_I + (p,q)$ on top of $R_I \cap (R_I)\mathrm{P} + (p,q)$ and ends at the bottom of a $R_{II} + (p+r-1, q+s-1)$ as shown in Figure 8.VIII.

As drawn in Figure 8.VI, the line segment aa$'''$ lies totally in $\overline{R_I}$ connecting its bottom horizontal boundary with its top: so (R_i)P has the property that it stripes across $R_I + (m,n)$ if and only if it passes through the interior of the line segment aa$'''+(m,n)$. But one must bear in mind that the picture does not represent all cases as c$'$ may not be located in the principal unit square so that the line segment aa$'''$ may not lie totally in $\overline{R_I}$. Nevertheless, because of the way (R_I)P begins and ends it obeys this property when $(m,n) = (0,0)$ and $(m,n) = (p-1,q)$. Consequently, it also obeys this property for at all intermediary lattice points. By *intermediary* we mean at (m,n) where $0 \le m \le p-1$ and $0 \le n \le q$. Similarly for (R_{II})P it satisfies the property for $(m,n) = (p,q)$, $(m,n) = (p+r-1, q+s-1)$, and thus for all intermediary (m,n).

The number of times $(R_i)P$ passes through a lattice translate aa$'''$ equals the number different lines $x =$ integer crossed by $(R_i)P$, which equals p for (R_I)P and r for (R_{II})P.

Similarly, a parallelogram (R_i)P intersects a lattice translate of R_{II} if and only if (R_i)P crosses the same a lattice translate of a$'''$a$''$; and the number of times this happens equals the number lines $y =$ integer crossed by (R_i)P, which equals q for (R_I)P and s for (R_{II})P.

Remark. A consequence of the fact that right contracting boundary of (R_I)P is contained in a lattice translate of ab is that $\phi R_I \cap R_{II} \neq \emptyset$ which means $q > 0$. But this we aready know from the property that P is hyperbolic. Similarly, $r > 0$. However, either p or s could be 0, but not both.

Step 5. *The transition matrix associated with mapping of the Markov generator $\mathcal{R}^*$ by ϕ coincides with matrix which specifies the automorphism–namely,* P*− the edge graph of which is illustrated in Figure 8.V.*

From what we have established there are $N^* = p + q + r + s$ parallelograms R_i^* in $\mathcal{R}^*$. We separate the subscripts into four sets:

$$\{1, \dots, p\} \tag{1}$$

$$\{p + 1, \dots, p + q\} \tag{2}$$

$$\{p + q + 1, \dots, p + q + r\} \tag{3}$$

$$\{p + q + r + 1, \dots, p + q + r + s\}. \tag{4}$$

We label the members of $\mathcal{R}^*$ accordingly:

$$\begin{aligned} R_1^* \cup \cdots \cup R_p^* &= R_I \cap \phi R_I, \\ R_{p+1}^* \cup \cdots \cup R_{p+q}^* &= R_{II} \cap \phi R_I, \\ R_{p+q+1}^* \cup \cdots \cup R_{p+q+r}^* &= R_I \cap \phi R_{II}, \\ R_{p+q+r+1}^* \cup \cdots \cup R_{p+q+r+s}^* &= R_{II} \cap \phi R_{II}. \end{aligned}$$

As we have said before: if $\phi h_i(\mathrm{p}) \cap R_k^* \neq \emptyset$ for one horizontal $h_i(\mathrm{p})$ of R_i, then $\phi h_i(\mathrm{q}) \cap R_k^* \neq \emptyset$ for any other horizontal $h_i(\mathrm{q})$ of R_i. Therefore, the image of each parallelogram R_k^* contained in an R_i intersects the same elements of $\mathcal{R}^*$. From this we get that $\phi R_k^* \cap R_l^* \neq \emptyset$, equivalently $R_k^* \cap \phi^{-1} R_l^* \neq \emptyset$, whenever either

$$k \in (1) \cup (3) \text{ and } l \in (1) \cup (2),$$

or

$$k \in (2) \cup (4) \text{ and } l \in (3) \cup (4).$$

<u>Case II:</u> $\lambda > 0, \mu < 0$.

Except for the lattice points of the unit square labelled by o's we redefine the other ones in Figure VI by translating them by $\rho \mathrm{v}_\mu$ with ρ, chosen so that

$$(\bar{\mathrm{d}})\mathrm{P} = \mathrm{a}.$$

For example, $\mathrm{a}_{\mathrm{new}} = \mathrm{a}_{\mathrm{old}} + \rho \mathrm{v}_\mu$. Whereas before the point a was the fixed point o, now it isn't. Figures 8.IX(A) and 8.IX(B) depict new principal Markov fundamental

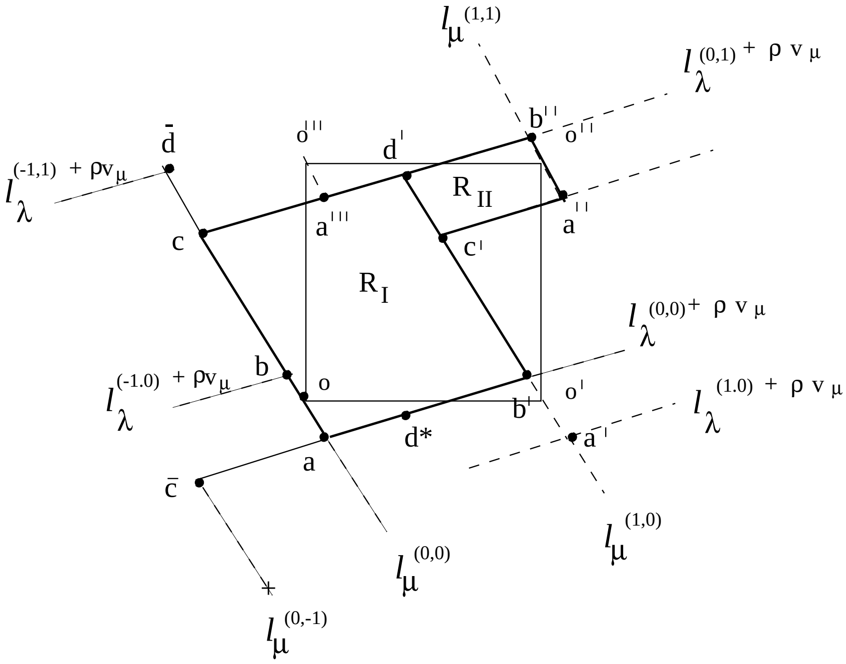


FIGURE 8.IX(A). New principal Markov fundamental region, a translation of the previous with o between a and b.

regions in which the old one of Figure VI has been translated by $\rho \mathrm{v}_\mu$. The difference in the two figures is the location of o with respect to b. In a moment we shall argue that Figure 8.IX(B) cannot occur.

STEP 1. *The new partition $\mathcal{R} = \{R_I, R_{II}\}$ is Markov.*

The proof of this is the same as in Step 1 of Case I except the verification of

$$\phi \partial_V \mathcal{R} \subset \partial \mathcal{R}_V \subset \partial \mathcal{R},$$
$$\phi^{-1} \partial_H \mathcal{R} \subset \partial_H \mathcal{R} \subset \partial \mathcal{R}$$

is more involved and is as follows.

For this new partition, (8.6) remains valid: namely,

$$\partial_H \mathcal{R} = \overline{\mathrm{c}}\mathrm{b}' \subset \ell_\lambda^a,$$

$$\partial_V \mathcal{R} = \mathrm{a}\overline{\mathrm{d}} \subset \ell_\mu.$$

Before there was a common fixed point on $\partial_H \mathcal{R}$ and $\partial_V \mathcal{R}$: namely, origin $=$ $\partial_H \mathcal{R} \cap \partial_V \mathcal{R}$, but now the intersection $\partial_H \mathcal{R} \cap \partial_V \mathcal{R}$ is no longer a fixed point. The line segment $\mathrm{a}\overline{\mathrm{d}}$ still contains the origin as a fixed point. The line segment $\overline{\mathrm{c}}\mathrm{b}$ also

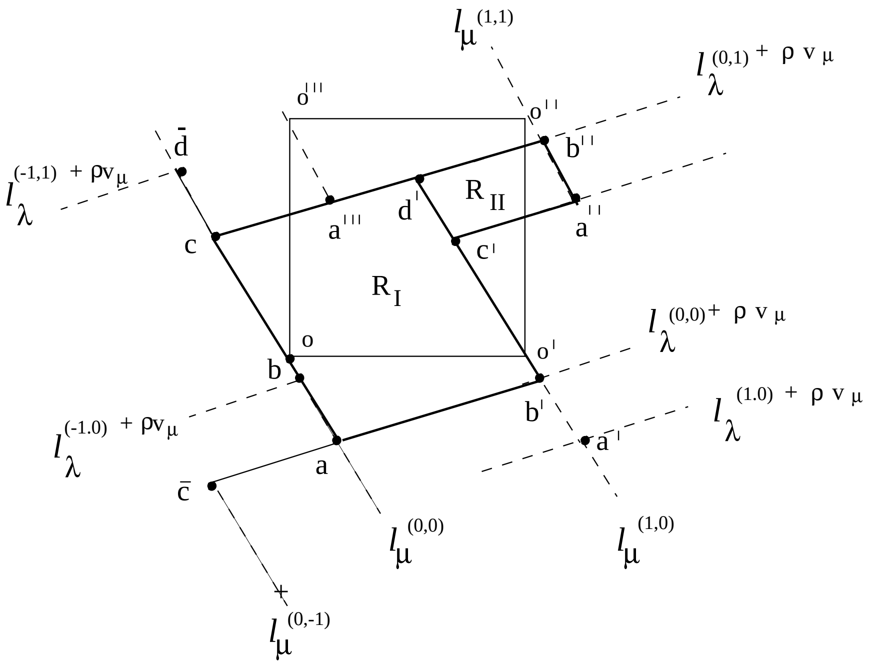


FIGURE 8.IX(B). New principal Markov fundamental region, a translation of the previous with o above b.

contains a fixed point, but this is not so obvious. This will be a consequence of the following version of (8.7):

$$
\begin{aligned}
&(\bar{\mathrm{c}}\mathrm{b}')\mathrm{P} \supset \text{lattice translate of } \bar{\mathrm{c}}\mathrm{b}' \\
&(\mathrm{a}\bar{\mathrm{d}})\mathrm{P} \subset \mathrm{a}\bar{\mathrm{d}};
\end{aligned}
\tag{8.11}
$$

Verification of the first relation is more difficult than before. That the second relation holds is immediate because the translation was chosen to make it so. The first relation rests on the property of the partition that $\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}$ are the only points on the contracting boundary which have equivalent ones on the expanding. There are no others!

Recall $\mathrm{d}^* = \mathrm{d}' - (0,1)$. Now because $(\bar{\mathrm{d}})\mathrm{P} = \mathrm{a}$, the point $\hat{\mathrm{a}} \equiv (\mathrm{d}^*)\mathrm{P}$ is a lattice translate of a and the line $\ell_\lambda^{\mathrm{a}}$ is mapped to the line $\ell_\lambda^{\hat{\mathrm{a}}}$ by P. See Figure 8.X. Let $\hat{\mathrm{c}}$, $\hat{\mathrm{b}}$, and $\hat{\mathrm{d}}$ be the points on the line $\ell_\lambda^{\hat{\mathrm{a}}}$ that are lattice translates of the points $\bar{\mathrm{c}}$, b', and d^*. We must show that $(\bar{\mathrm{c}}\mathrm{b}')\mathrm{P}$ contains $\hat{\mathrm{c}}\hat{\mathrm{b}}$. Because the point $(\mathrm{d}^*)\mathrm{P}$ lies on the line $\ell_\lambda^{\hat{\mathrm{a}}}$, so do the points $(\bar{\mathrm{c}})\mathrm{P}$ and $(\mathrm{b}')\mathrm{P}$. Since $(\mathrm{b})\mathrm{P}$ and $(\mathrm{c})\mathrm{P}$ are contained in the segment $\mathrm{a}\bar{\mathrm{d}}$ and the pair of points $(\mathrm{b}')\mathrm{P}$ and $(\bar{\mathrm{c}})\mathrm{P}$ are lattice translates of neither a, b, c, nor $\bar{d}$, this pair lies outside the segment $\hat{\mathrm{b}}\hat{\mathrm{c}}$. Since $(\mathrm{d}^*)\mathrm{P}$ lies inside, the segment, $(\bar{\mathrm{c}}\mathrm{b}')\mathrm{P}$ overlaps $\hat{\mathrm{c}}\hat{\mathrm{b}}$.

Now we are are in a position to prove that Figure8.IX(B) cannot occur. Let $\tilde{\mathrm{a}} = (p,q) + \rho \mathrm{v}_\mu$. Since $(\bar{\mathrm{d}})\mathrm{P} = \mathrm{a}$, we have that $\tilde{\mathrm{a}} = \mathrm{d}'\mathrm{P}$. Let $\tilde{\mathrm{b}}$ be the point equivalent

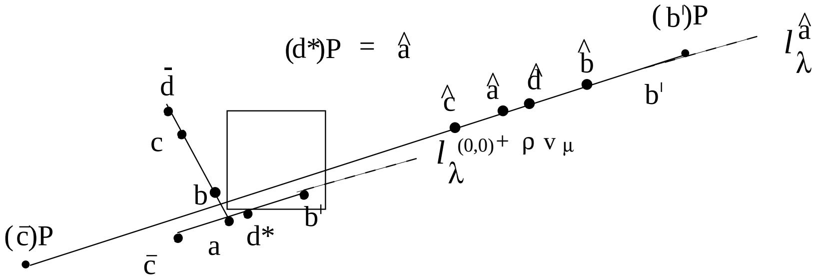


FIGURE 8.X. Image of expanding boundary of new principal Markov fundamental region.

to b′ on the line $\ell_\mu^{(p,q)}$, upon which ã lies; and let c̃ be the lattice translate of $\bar{c}$ on the line $\ell_\lambda^{(p,q)+\rho v_\mu}$. The segment $\tilde{\mathrm{c}}\tilde{\mathrm{b}}$ is a lattice translate of the horizontal boundary of $\overline{R_I} \cup \overline{R_{II}}$. If b were below o on the line $\ell_\mu^{(0.0)}$, then by virtue of the fact that μ is negative the point b′P would lie above (p,q) on this line, and hence above b̃. Since d′P = ã lies below b̃, the parallelogram R_IP would straddle the part of its own horizontal boundary: namely the segment $\tilde{\mathrm{c}}\tilde{\mathrm{b}}$. This would contradict the fact that R_IP is a subset of a fundamental region.

STEP 2. *The family $\mathcal{R}^* = \{R_k^* : 1 \le k \le N^*\}$ consisting of all connected components of the sets $R_i \cap \phi^{-1} R_j \in \mathcal{R} \vee \phi^{-1}\mathcal{R}$ is Markov partition.*

Proof is the same as Step 2 of Case I.

STEP 3. *The Markov partition $\mathcal{R}^*$ is a generator.*

Proof, same as Case I, Step 3.

STEP 4. *Let $\#(\phi R_i \cap R_j)$ denote the number of disjoint parallelograms in the intersection $\phi R_i \cap R_j$.*

$$\begin{aligned} \#(\phi R_i \cap R_I) &\equiv \textit{ number of lines } x=0, x=1, \ldots, \textit{ traversed by } (R_i)P; \\ \#(\phi R_i \cap R_{II}) &\equiv \textit{ number of lines } y=1, y=2, \ldots, \textit{ traversed by } (R_i)P, \end{aligned}$$

which leads to

$$\begin{aligned} \#(\phi R_I \cap R_I) &= x\textit{-coordinate of } (0,1)\mathrm{P} &&= p; \\ \#(\phi R_{II} \cap R_I) &= x\textit{-coordinate of } (1,1)\mathrm{P} - (1,0)\mathrm{P} &&= q; \\ \#(\phi R_I \cap R_{II}) &= y\textit{-coordinate of } (0,1)\mathrm{P} &&= r; \\ \#(\phi R_{II} \cap R_{II}) &= y\textit{-coordinate of } (1,1)\mathrm{P} - (1,0)\mathrm{P} &&= s. \end{aligned}$$

Argument similar to Case I, Step 4 but based on Figure 8.XI. There is a slight difference here. The vertical line segment aa′′′ is offset from the line $x = 0$, and

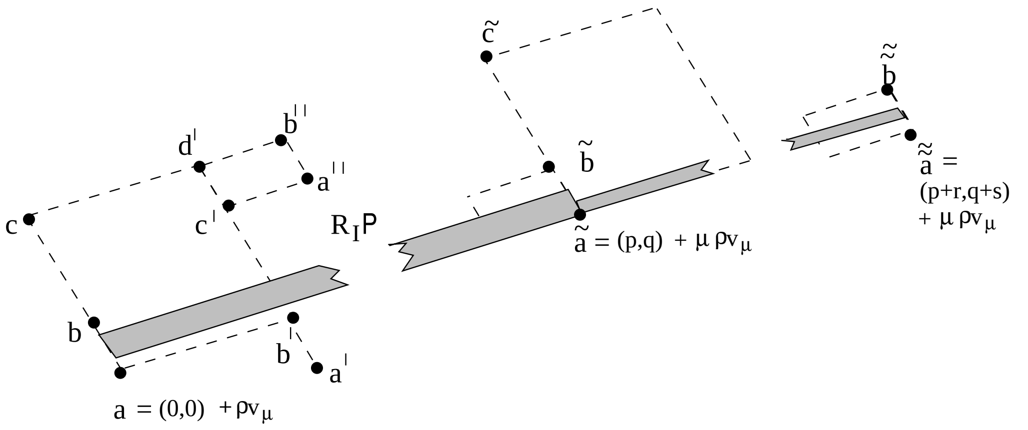


FIGURE 8.XI. How the new (R_I)P and (R_{II})P intersect various fundamental regions

a$'''$a$''$ does not lie on $y = 1$. If $\rho v_\mu = (\xi, \eta)$, then

$$\#(\phi R_i \cap R_I) \equiv \text{ number of lines } x - \xi = 0, x - \xi = 1, \dots, \text{ traversed by } (R_i)P;$$
$$\#(\phi R_i \cap R_{II}) \equiv \text{ number of lines } y - \eta = 1, y - \eta = 2, \dots, \text{ traversed by } (R_i)P,$$

To get the desired result we observe that the parallelgram $(R_i)P$ traverses the line $x - \xi = m$ if and only if it traverses the line $x = m$. Similarly the parallelogram traverses $y - \eta = m$ if and only if it traverses $y = m$.

STEP 5. *The transition matrix associated with mapping of the Markov generator $\mathcal{R}^*$ by ϕ coincides with matrix which specifies the automorphism–namely,* P*– the edge graph of which is illustrated in Figure 8.V.*

Argument same as Case I Step 5.

Case III: $\lambda < 0, \mu > 0$.
The sets R_i are the same as for Case I.

STEP 1. *The new partition $\mathcal{R} = \{R_I, R_{II}\}$ is Markov.*

This depends on establishing (8.6): namely

$$\overline{\mathrm{c}}\mathrm{b}' \subset (\overline{\mathrm{c}}\mathrm{b}')P$$
$$(\mathrm{a}\overline{\mathrm{d}})P \subset \mathrm{a}\overline{\mathrm{d}}.$$

The argument for the first relation is similar to that of Case II, Step 1. The argument for the second is the same as (8.7) of Case I.

STEP 2. *The family $\mathcal{R}^* = \{R_k^* : 1 \leq k \leq N^*\}$ consisting of all connected components of the sets $R_i \cap \phi^{-1} R_j \in \mathcal{R} \vee \phi^{-1}\mathcal{R}$ is Markov partition.*

Proof is the same as Step 2 of Case I.

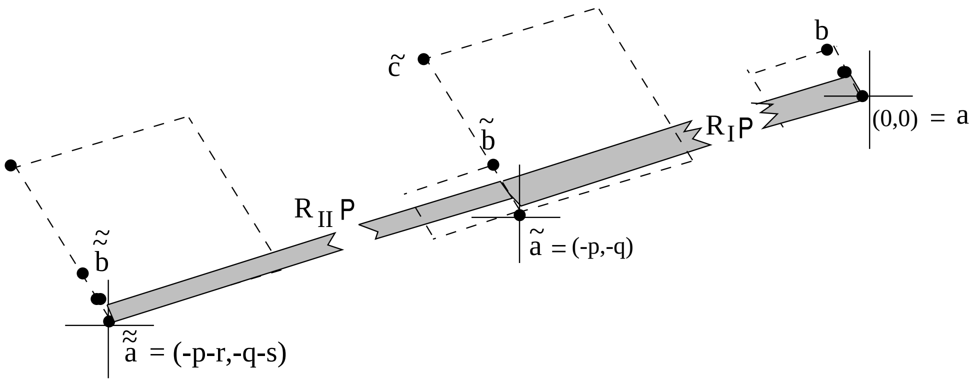


FIGURE 8.XII. How (R_I)P and (R_{II})P intersect various fundamental regions for Case III.

STEP 3. *The Markov partition $\mathcal{R}^*$ is a generator.*

Proof, same as Case I, Step 3.

STEP 4. *Let $\#(\phi R_i \cap R_j)$ denote the number of disjoint parallelograms in the intersection $\phi R_i \cap R_j$.*

$$\#(\phi R_i \cap R_I) \equiv \text{ number of lines } x = 0, x = -1, \ldots, \text{ traversed by } (R_i)P;$$
$$\#(\phi R_i \cap R_{II}) \equiv \text{ number of lines } y = -1, y = -2, \ldots, \text{ traversed by } (R_i)P,$$

which leads to

$$\begin{aligned}
\#(\phi R_I \cap R_I) &= x\text{-coordinate of } (0,1)\mathrm{P} &&= -p;\\
\#(\phi R_{II} \cap R_I) &= x\text{-coordinate of } (1,1)\mathrm{P} - (1,0)\mathrm{P} &&= -q;\\
\#(\phi R_I \cap R_{II}) &= y\text{-coordinate of } (0,1)\mathrm{P} &&= -r;\\
\#(\phi R_{II} \cap R_{II}) &= y\text{-coordinate of } (1,1)\mathrm{P} - (1,0)\mathrm{P} &&= -s.
\end{aligned}$$

Argument the same as Case I, step 4 but based on Figure 8.XII.

STEP 5. *The transition matrix associated with mapping of the Markov generator $\mathcal{R}^*$ by ϕ coincides with matrix which specifies the automorphism–namely,* P*– the edge graph of which is illustrated in Figure 8.V.*

Argument same as Case I Step 5.

<u>Case IV:</u> $\lambda < 0, \mu < 0$.
The sets R_i are the same as in Case II– *i.e.*, translations of the ones of Case I.

STEP 1. *The new partition $\mathcal{R} = \{R_I, R_{II}\}$ is Markov.*

The argument is the same as in Case II, Step1

STEP 2. *The family $\mathcal{R}^* = \{R_k^* : 1 \leq k \leq N^*\}$ consisting of all connected components of the sets $R_i \cap \phi^{-1}R_j \in \mathcal{R} \vee \phi^{-1}\mathcal{R}$ is Markov partition.*

Proof same as Case I, Step 2.

STEP 3. *The Markov partition $\mathcal{R}^*$ is a generator.*

Proof, same as Case I, Step 3.

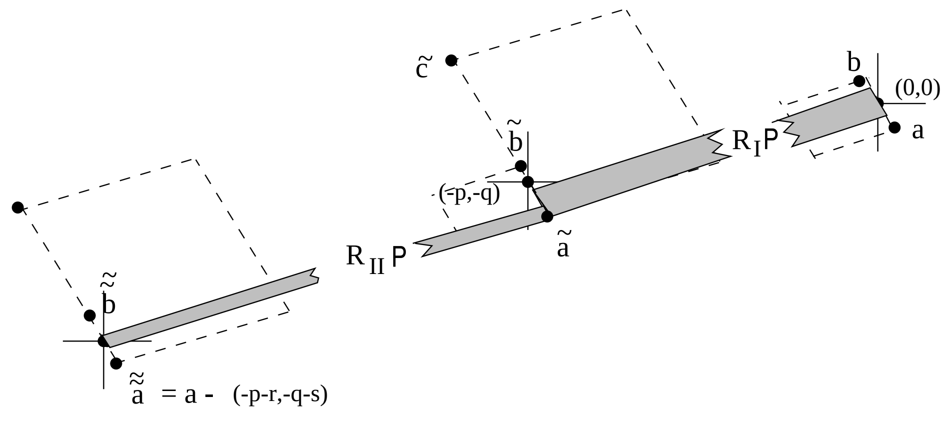


FIGURE 8.XIII. How (R_I)P and (R_{II})P intersect various fundamental regions Case IV

STEP 4. *Let $\#(\phi R_i \cap R_j)$ denote the number of disjoint parallelograms in the intersection $\phi R_i \cap R_j$.*

$$\#(\phi R_i \cap R_I) \equiv \text{ number of lines } x = 0, x = -1, \ldots, \text{ traversed by } (R_i)P;$$
$$\#(\phi R_i \cap R_{II}) \equiv \text{ number of lines } y = -1, y = -2, \ldots, \text{ traversed by } (R_i)P,$$

which leads to

$$\begin{aligned}
\#(\phi R_I \cap R_I) &= x\text{-coordinate of } (0,1)\mathrm{P} &&= -p;\\
\#(\phi R_{II} \cap R_I) &= x\text{-coordinate of } (1,1)\mathrm{P} - (1,0)\mathrm{P} &&= -q;\\
\#(\phi R_I \cap R_{II}) &= y\text{-coordinate of } (0,1)\mathrm{P} &&= -r;\\
\#(\phi R_{II} \cap R_{II}) &= y\text{-coordinate of } (1,1)\mathrm{P} - (1,0)\mathrm{P} &&= -s.
\end{aligned}$$

Argument the same as Case II, Step 4 but based on Figure 8.XIII.

STEP 5. *The transition matrix associated with mapping of the Markov generator $\mathcal{R}^*$ by ϕ coincides with matrix which specifies the automorphism–namely,* P*– the edge graph of which is illustrated in Figure 8.V.*

Argument same as Case I Step 5.

□

Exercises.

8.1 Let (X, ϕ) be the dynamical system $\mathcal{R} = \{R_i \ : i = 1, 2, 3\}$ the Markov partition of Example 3.3. Show the partition $\mathcal{S} = \{S_i \ : i = 0, 1\}$, where $S_0 = R_1 \cup R_2$ and $S_1 = R_3$, is Markov. Hint: $\mathcal{S}^{(2)} = \mathcal{R}$. Observe that associated directed graph G is vertex labelled graph of Figure 2.II.

8.2 What is the maximum number of pre-images of π in the last example? In Theorem 8.4?

9. Epilogue. The main theorem of Section 6 and its converse exhibit a duality between factor maps and Markov partitions. We are thus presented with a type of "chicken verses egg" question: which are more fundamental for getting concrete symbolic representations for concrete dynamical systems, Markov partitions or factor maps? In Example 3.2, we have with equal ease defined a factor map arithmetically producing from it a Markov partition and constructed a Markov partition thereby obtaining a factor map. However,in Section 8, we only constructed a Markov partition for an arbitrary hyperbolic automorphism of the two-dimensional torus. In trying to repeat the success of dimension two by drawing Markov partitions for hyperbolic toral automorphisms in dimension three, one is doomed to failure as Bowen proves in [Bo2]. He shows that no point on the contracting boundary of a member of a Markov partition can there exist a tangent plane. The boundary is a forced to be a fractal. Even if such a figure could be constructed, rendering is certainly difficult, let alone incorporating it in a tiling of three space in a manner suitable for viewing. And then what about four dimensions and higher?. Therefore, in any dimension other than two an arithmetic method seems the only hope.

Kenyon, Vershik [KV], and Praggastis [P] attack the problem in this manner and obtain factor maps arithmetically for hyperbolic toral automorphisms.

Ultimately the answer to our chicken-egg question will probably turn out to be that neither takes precedence over the other. Bowen's proof of existence of Markov partitions and hence finite factor maps for general hyperbolic axiom A diffeomorphisms supports this point of view [Bo1]. In this work Bowen uses an argument employing a combination of both methods. and bootstraps his way to the desired result. It starts with a cover whose members have small diameter with respect to the expansive constant. A topological Markov shift is constructed by using the labels of the members as symbols and defining transitions according to the rule that the j-th symbol follows the i-th if the image of the i-th member of the cover intersects the j-th. Using stable and unstable manifolds one is able to define a factor map from the symbolic shift to the phase space. This symbolic extension is much too big. The cardinality of a pre-images under this map can be non-denumerable. However, the images of cylinder sets form a second cover whose members are abstract rectangles obeying a Markovian property. This cover though may fail to form a topological partition for two reasons: some members may have no interior; and pairs of them may overlap in more than just boundary points. This cover is pared down by eliminating those sets with no interior. The overlap problem is solved by using some geometry of abstract rectangles: namely two abstract rectangles overlapping in an open set can be partitioned into nonoverlapping abstract rectangles. The last step in getting a topological partition is to partition the members of this last cover abstract rectangles, no pair of which overlaps in an open set. Through all of this Markov property is still maintained by the members of this final partition. That this is a generator is gotten by the same argument as our Propostion 5.8.

Returning to the line of investigation of Kenyon, Vershik, and Praggastis, there remain things to be understood, such as the connectivity of cylinder set images under arithmetically defined factor maps. Furthermore, the general case of Markov partitions for hyperbolic automorphims of n-dimensional tori has not yet been treated.

References

[ATW] Roy Adler, Charles Tresser, Patrick A. Worfolk, *Topological conjugacy of linear endomorphisms of the 2-torus*, Trans. Amer. Math. Soc. ((to appear)).

[AW] Roy L. Adler and Benjamin Weiss, *Similarity of automorphisms of the torus*, Memoirs American Math. Soc. **98** (1970).

[Be] K. Berg, *On the conjugacy problem for K-systems* (1967), University of Minnesota.

[Bo1] R. Bowen, *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms,* Lecture Notes in Mathematics, Vol. 470, Springer-Verlag, Berlin, Heidelberg, New York, 1975.

[Bo2] R. Bowen, *Markov partitions are not smooth*, Proc. Amer. Math. Soc. **71** (1970), 130-132.

[KV] R. Kenyon and A. Vershik, *Arithmetic construction of sofic partitions of hyperbolic toral automorphisms*, Ecole Normale Superieure de Lyon.

[LM] D. Lind and B Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge University Press, Cambridge, New York, Melbourne, 1995.

[P1] B Praggastis, *Markov partitions for hyperbolic toral automorphisms* (1992), University of Washington.

[W] R. F. Williams, *The "DA" maps of Smale and structural stability*, Proc. Symp. in Pure Math. **14** (1970), AMS, 329-334.