1. Examples of dynamical systems

§1.1 Course outline

This is a 10-lecture course on ergodic theory. Each lecture will be given by slides (which will also be available via the course homepage on the MAGIC website). Each lecture will also have an associated handout, containing more information about the material or supplying details that were only sketched in the lectures. Each handout also contains a number of exercises, and you are strongly recommended to try doing these.

§1.2 Introduction

Ergodic theory is a branch of dynamical systems. A dynamical system consists of a space X (the *state space* or *phase space*) and a rule that governs how points in X evolve over time. Time can vary either discretely or continuously. In the case of discrete time the dynamics is governed by iterating a map $T : X \to X$. For $n \geq 0$ we write $T^n = T \circ \cdots \circ T$ (n times) for the nth iterate of T. If T is invertible, then this definition makes sense for all $n \in \mathbb{Z}$. In the case of continuous time, the dynamics is normally governed by a differential equation and this gives rise to a *flow*: a continuous one-parameter family T_t of transformations of X with the property that $T_{t+s}(x) = T_t(T_s(x))$. Again, this may make sense just for $t \in \mathbb{R}^+$ (in which case we call T_t a *semi-flow*) or for all $t \in \mathbb{R}$.

Throughout this course we will almost always work in the context of discrete time dynamical systems. However, almost everything we discuss has a natural analogue in the case of continuous time.

Given $x \in X$ (the *initial condition*), the set $\{T^n(x) \mid n \in \mathbb{Z}\}$ is called the *orbit* or *trajectory* of x. If we restrict to positive n, then we call this the *forward orbit* of x.

Some orbits may be periodic (that is: $T^n x = x$ for some $n \geq 1$), whereas other orbits may fill out whole regions of the phase space. Suppose that X has some metric structure (for example X could be a subset of $\mathbb{R}^n$). Further, suppose that the dynamical system is continuous. Then if $x, y \in X$ are nearby, then by continuity, $T(x), T(y)$ will also be reasonably close. However, as we keep iterating T the two orbits may eventually diverge and have very different behaviours. This is called *sensitive dependence on initial conditions* (or *chaos* in popular culture). This suggests that, in general, studying individual orbits is a hard problem. Ergodic theory takes a more qualitative view: we aim to study the long term behaviour of typical orbits.

Consider the following example of a continuous time dynamical system, namely the geodesic flow on a torus. Let $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$ denote the two-dimensional torus. Let X denote the unit tangent bundle of $\mathbb{T}^2$. The unit tangent bundle of $\mathbb{T}^2$ consists of the set of all pairs of the form $x = (p, v)$ where p is a point $p \in \mathbb{T}^2$ and v is a unit vector at p. Thus the unit tangent bundle is the set of all points and all possible directions through each point.

Given $x = (p, v) \in X$, there is a unique straight line in $\mathbb{T}^2$ through the point p in the direction v. We can extend this line infinitely in both directions by using the identifications. The continuous time dynamical system T_t acts as follows: consider the point $x = (p, v)$ and consider the unique straight line through p in the direction v, then $T_t p$ is the point in X determined by travelling along the this straight line at constant speed 1 for time t. See Figure 1.1.

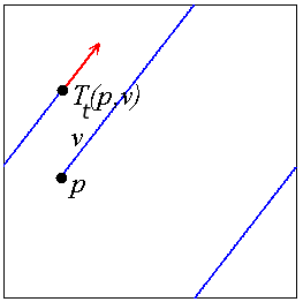


Figure 1.1: The point $x = (p, v)$ and its image at time t

One can easily check that some orbits are periodic whereas some orbits in X project to dense sets in $\mathbb{T}^2$; see Figure 1.2. One can ask: for which directions v does the orbit of $x = (p, v)$ project to a dense subset of $\mathbb{T}^2$? More specifically: when is the projection of this orbit 'uniformly distributed' in $\mathbb{T}^2$? In other words: is the proportion of time that the projection to $\mathbb{T}^2$ of the orbit of (p, v) spends in a given set $A \subset \mathbb{T}^2$ equal to the relative size of A? This certainly seems to be the case in Figure 1.2: no region of $\mathbb{T}^2$ appears to be favoured over any other region. Ergodic theory addresses questions like this.

Now let $T : X \to X$ be an arbitrary transformation of an arbitrary phase space. In order to study 'typical' orbits, and indeed to define 'typical', we need to use measure theory. We equip X with the structure of a measure space and in particular we assume that we have a finite measure μ (so that $\mu(X) < \infty$). By saying that a point is 'typical' we will normally mean that it belongs to a set of full μ-measure.

Given a subset $A \subset X$ we can ask, analogously to the above, the frequency with which the orbit of a point $x \in X$ lies in A. This frequency is

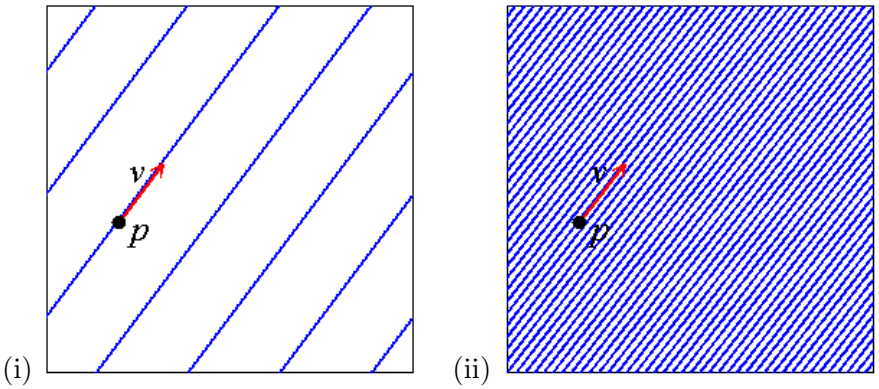


Figure 1.2: (i) A periodic orbit, and (ii) an orbit whose projection to $\mathbb{T}^2$ is dense

easily seen to be given by

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \chi_A(T^j x).$$

A basic question in ergodic theory is to understand such limits. The most basic question is whether the above limit exists for a reasonably large set of x, and whether this limit (thought of as the 'time average' of the orbit of x is equal to the proportion of the set X that is occupied by A (the 'space average'). That is, when is it true that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \chi_A(T^j x) = \frac{\mu(A)}{\mu(X)}$$

for a large set of points x? The correct hypothesis to make is that T is *ergodic* with respect to the measure μ, a property that we shall discuss in some detail in future lectures.

More generally, suppose we have a function $f : X \to \mathbb{R}$. We can think of evaluating f at a point $x \in X$ as taking a measurement or observation. Given a point $x \in X$ and regarding its orbit under a dynamical system T as how x evolves in time, we can make the sequence of observations $f(x), f(Tx), \ldots, f(T^n x), \ldots$ and the take the average value:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(T^j x).$$

It is natural to expect that, under suitable hypotheses, that this limit is equal to the integral of f. Indeed, this forms the content of the following theorem.

Theorem 1.1 (Birkhoff's Ergodic Theorem (1931))
Let T be an ergodic measure-preserving transformation of a probability space $(X, \mathcal{B}, \mu)$. Let $f \in L^1(X, \mathcal{B}, \mu)$. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(T^j x) = \int f \, d\mu$$

for μ-almost every $x \in X$.

(We define all the terms in the above theorem in a future lecture.)

Birkhoff's Ergodic Theorem can be seen as a version of the strong law of large numbers: long-term observed behaviour is equal to the expected value. Indeed, there any many connections between ergodic theory and probability theory, some of which we will discuss in future lectures.

The remainder of this lecture is devoted to describing a variety of dynamical systems that will be useful throughout the course.

§1.3 Maps on the circle

There are several dynamical systems of interest that are defined on the unit circle K. Recall the the circle is also a group. The circle can be written either additively as $\mathbb{R}/\mathbb{Z}$ (and we will often abuse notation by writing an element of $\mathbb{R}/\mathbb{Z}$ as $x \in [0, 1)$ with the understanding that 0 and 1 are identified), a multiplicatively as

$$S^1 = \{z \in \mathbb{C} \mid |z| = 1\}.$$

Usually we will use additive notation. Moreover, we will often abuse notation by writing identifying a point $x \in \mathbb{R}$ with its coset $x + \mathbb{Z} \in \mathbb{R}/\mathbb{Z}$. We will also write $x + \mathbb{Z}$ as $x \bmod 1$.

More generally, we will be interested in the d-dimensional torus K^d, and this is again a group. This can also be written either additively as $\mathbb{R}^d/\mathbb{Z}^d$, or multiplicatively as the d-fold product $S^1 \times \cdots S^1$. Again we usually use additive notation and make a similar abuse of notation as above.

§1.3.1 Rotations

Fix $\alpha \in [0, 1)$. Define the map $T_\alpha : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ by

$$T_\alpha(x + \mathbb{Z}) = x + \alpha + \mathbb{Z} \bmod 1.$$

(We often abuse this notation and instead write $T_\alpha(x) = x + \alpha \bmod 1$.) We call T_α a *rotation by α* on the circle. Multiplicatively this would be the map

$$z \mapsto e^{2\pi i\alpha} z$$

and this rotates the unit circle through an angle of $2\pi\alpha$.

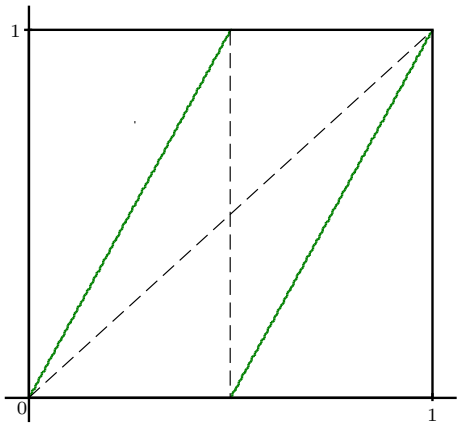


Figure 1.3: The graph of the doubling map

Clearly, if α is rational then every point of T_α is periodic. Indeed, if $\alpha = p/q$ with $p, q \in \mathbb{Z}$ and p, q coprime, then T_α^q is the identity. Conversely, if α is irrational then it is simple to observe that there are no periodic orbits. Indeed, the following is true:

Proposition 1.2
Define $T_\alpha : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ *by* $T_\alpha(x) = x + \alpha \bmod 1$.

(i) If α *is rational then every point* $x \in \mathbb{R}/\mathbb{Z}$ *is periodic and has the same period.*

(ii) If α *is irrational then the orbit of every point* $x \in \mathbb{R}/\mathbb{Z}$ *is dense.*

More generally, we can define a rotation of any group. Let G be a group and fix an element $g \in G$. Define the *(left) rotation by* g to be the map

$$T : G \to G : x \mapsto gx.$$

§1.3.2 The doubling map

Define the map $T : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ by $T(x) = 2x \bmod 1$. This is the *doubling map*. (Multiplicatively this is the map $z \mapsto z^2$.) See Figure 1.3.

The doubling map is a classic example of a chaotic map. According to Devaney's definition of chaos, a chaotic map must have:

(i) sensitive dependence on initial conditions; (A map $T : X \to X$ has sensitive dependence on initial conditions if: there exists an $\varepsilon > 0$ such that for all $x \in X$ and all $\delta > 0$ there exists $y \in X$ and $n \geq 0$ such that $d(x, y) < \delta$ but $d(T^n x, T^n y) \geq \varepsilon$. A related, and perhaps more important (in the context of topological dynamics) is *expansivity*: T

is expansive if there exists $\varepsilon > 0$ such that $d(T^n x, T^n y) < \varepsilon$ for all n implies $x = y$.)

(ii) the set of periodic points is dense;

(iii) a dense orbit.

(Note that this definition of chaos is topological in nature and, therefore, not strictly speaking an ergodic theoretic concept. Also note that there are many dynamical systems of interest in ergodic theory that do not satisfy the above definition of chaos; an example is an irrational rotation of a circle.)

The doubling map is closely related to number theory. Any $x \in [0,1]$ has a binary expansion of the form

$$x = \frac{x_0}{2} + \frac{x_1}{2^2} + \cdots + \frac{x_n}{2^{n+1}} + \cdots = \sum_{j=0}^{\infty} \frac{x_j}{2^{j+1}}$$

where $x_j \in \{0,1\}$. We can use the doubling map to calculate the digits x_j by noting that $x_n = 0$ if $T^n \in [0,1/2)$ and $x_n = 1$ if $T^n(x) \in [1/2,1]$.

Let $\Sigma = \{(x_j)_{j=0}^{\infty} \mid x_j \in \{0,1\}\}$ denote the space of all possible infinite sequences of 0s and 1s. Define the shift map $\sigma : \Sigma \to \Sigma$ by

$$\sigma(x_0, x_1, x_2, \ldots) = (x_1, x_2, \ldots).$$

That is, σ takes a sequence, deletes the digit in the 0th place and then shifts the sequence one place to the left.

Define a projection map $\pi : \Sigma \to \mathbb{R}/\mathbb{Z}$ by

$$\pi(x_0, x_1, x_2, \ldots) = \sum_{n=0}^{\infty} \frac{x_n}{2^{n+1}}.$$

Then

$$\begin{aligned} T(\pi(x_0, x_1, x_2, \ldots) &= T\left(\sum_{n=0}^{\infty} \frac{x_n}{2^{n+1}}\right) \\ &= x_0 + \sum_{n=1}^{\infty} \frac{x_n}{2^n} \bmod 1 \\ &= \sum_{n=1}^{\infty} \frac{x_n}{2^n} \end{aligned}$$

as $x_0 = 0$ or 1; moreover,

$$\begin{aligned} \pi(\sigma(x_0, x_1, x_2, \ldots)) &= \pi(x_1, x_2, \ldots) \\ &= \sum_{n=1}^{\infty} \frac{x_n}{2^n}. \end{aligned}$$

Thus $T\pi = \pi\sigma$.

The map π codes the dynamics of T symbolically and can be used to prove many dynamical properties of T. In future lectures we will use symbolic dynamics to study a variety of dynamic and ergodic-theoretic properties of a wide class of dynamical systems.

Note that π enjoys the following properties:

(i) π is surjective,

(ii) π is not injective (points $x \in \mathbb{R}/\mathbb{Z}$ with a base 2 expansion ending in all 0s have another base 2 expansion ending in all 1s; this is precisely analogous to the fact that decimal expansions are not unique if they end in all 0s or all 9s); however, it is injective on the complement of this set.

We discuss symbolic dynamical systems in more detail below. Let us record here that symbolic dynamics can be used to prove the following proposition.

Proposition 1.3
The doubling map is chaotic.

Constructing symbolic dynamics for the doubling map is particularly easy as there is a natural coding using base 2 expansions. More generally, we can view this in the following way. Write $\mathbb{R}/\mathbb{Z} = I_0 \cup I_1$ where $I_0 = [0, 1/2]$ and $I_1 = [1/2, 1]$. Given $x \in \mathbb{R}/\mathbb{Z}$ we can determine the coding $(x_0, x_1, \ldots)$ by noting that $T^n x \in I_{x_n}$. That is, we get the symbolic coding for x by recording the sequence of elements of the partition I_0, I_1 that the orbit of x visits. Note that x has a unique coding if and only if the orbit of x never lands on the end-points of I_0 or I_1. This idea of partitioning a set into a finite number of regions and coding the orbit of a point x by recording the elements of this partition that the orbit hits is a particularly powerful tool in the study of dynamical systems, particularly those with hyperbolic behaviour.

§1.4 Endomorphisms and automorphisms of the torus

Let $X = \mathbb{R}^k/\mathbb{Z}^k$ be the k-dimensional torus. Let $A = (a_{ij})$ be a $k \times k$ matrix with entries in $\mathbb{Z}$ and with $\det A \neq 0$. We can define a linear map $\mathbb{R}^k \to \mathbb{R}^k$ by

$$\begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} \mapsto A \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}.$$

For brevity, we shall often abuse this notation by writing this as $(x_1, \ldots, x_k) \mapsto A(x_1, \ldots, x_k)$.

Since A is an integer matrix it maps $\mathbb{Z}^k$ to itself. We claim that A allows us to define a map

$$T = T_A : X \to X : (x_1, \ldots, x_k) + \mathbb{Z}^k \mapsto A(x_1, \ldots, x_k) + \mathbb{Z}^k.$$

Again, we shall often abuse notation and we will often write $T(x_1, \ldots, x_k) = A(x_1, \ldots, x_k) \bmod 1$.

To see that this map is well defined, we need to check that if $x + \mathbb{Z}^k = y + \mathbb{Z}^k$ then $Ax + \mathbb{Z}^k = Ay + \mathbb{Z}^k$. But this is clear: if $x, y \in \mathbb{R}^k$ give the same point in the torus, then $x = y + n$ for some $n \in \mathbb{Z}^k$. Hence $Ax = A(y + n) = Ay + An$. As A maps $\mathbb{Z}^k$ to itself, we see that $An \in \mathbb{Z}^k$ so that Ax, Ay determine the same point in the torus.

Definition. Let $A = (a_{ij})$ denote a $k \times k$ matrix with integer entries such that $\det A \neq 0$. Then we call the map $T_A : \mathbb{R}^k/\mathbb{Z}^k \to \mathbb{R}^k/\mathbb{Z}^k$ a *linear toral endomorphism*.

The map T is not invertible in general. However, if $\det A = \pm 1$ then A^{-1} exists and is an integer matrix. Hence we have a map T^{-1} given by

$$T^{-1}(x_1, \ldots, x_k) = A^{-1}(x_1, \ldots, x_k) \bmod 1.$$

One can easily check that T^{-1} is the inverse of T.

Definition. Let $A = (a_{ij})$ denote a $k \times k$ matrix with integer entries such that $\det A = \pm 1$. Then we call the map $T_A : \mathbb{R}^k/\mathbb{Z}^k \to \mathbb{R}^k/\mathbb{Z}^k$ a *linear toral automorphism*.

Remark The reason for this nomenclature is clear. If T_A is either a linear toral endomorphism or linear toral automorphism, then it is an endomorphism or automorphism, respectively, of the torus regarded as an additive group.

Example. Take A to be the matrix

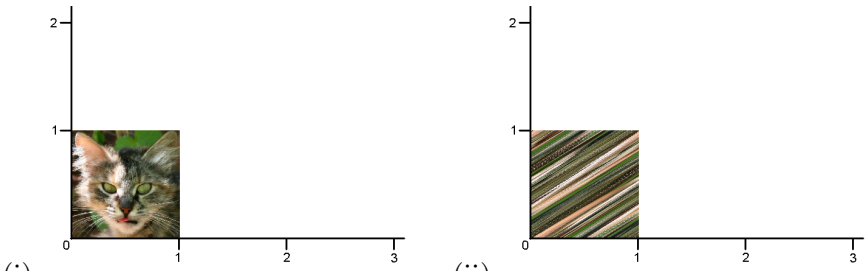
$$A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

and define $T : \mathbb{R}^2/\mathbb{Z}^2 \to \mathbb{R}^2/\mathbb{Z}^2$ to be the induced map:

$$T(x_1, x_2) = (2x_1 + x_2 \bmod 1, x_1 + x_2 \bmod 1).$$

Then T is a linear toral automorphism and is called Arnold's Cat map (CAT stands for 'C'ontinuous 'A'utomorphism of the 'T'orus). See Figure 1.4.

Definition. Suppose that $\det A = \pm 1$. Then we call T a *hyperbolic* toral automorphism if A has no eigenvalues of modulus 1.



(i) (ii)

Figure 1.4: (i) The torus, and (ii) its image under the Cat map

§1.5 The Gauss map

Every $x \in (0,1)$ can be expressed as a continued fraction:

$$x = \cfrac{1}{x_0 + \cfrac{1}{x_1 + \cfrac{1}{x_2 + \cfrac{1}{x_3 + \cdots}}}} \tag{1.1}$$

for $x_n \in \mathbb{N}$.

For example,

$$\begin{aligned} \frac{-1+\sqrt{5}}{2} &= \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1+\cdots}}}}, \\ \pi &= 3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1 + \cfrac{1}{292+\cdots}}}}. \end{aligned}$$

One can show that rational numbers have a *finite* continued fraction expansion (that is, the above expression terminates at x_n for some n). Conversely, it is clear that a finite continued fraction expansion gives rise to a rational number.

Thus each irrational $x \in (0,1)$ has an infinite continued fraction expansion of the form (1.1). Moreover, one can show that this expansion is unique. For brevity, we will often write (1.1) as $x = [x_0, x_1, x_2, \ldots]$.

Recall that the doubling map $x \mapsto 2x \bmod 1$ can be used to determine the base 2 expansion of x. We introduce a dynamical system that allows us to determine the continued fraction expansion of x.

We can read off the numbers x_i from the transformation $T : [0,1) \to [0,1)$ defined by $T(0) = 0$ and, for $0 < x < 1$,

$$T(x) = \begin{cases} 0 & \text{if } x = 0, \\ \left\{\frac{1}{x}\right\} = \frac{1}{x} \bmod 1 & \text{if } 0 < x < 1. \end{cases}$$

See Figure 1.5. Then

$$x_0 = \left[\frac{1}{x}\right],\ x_1 = \left[\frac{1}{Tx}\right], \ldots, x_n = \left[\frac{1}{T^n x}\right].$$

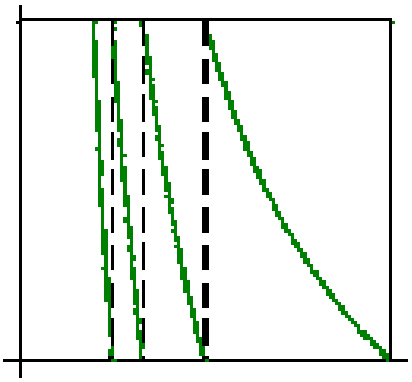


Figure 1.5: The graph of the Gauss map

§1.6 Shifts of finite type

Let $S = \{1, 2, \ldots, k\}$ be a finite set of symbols. We will be interested in sets consisting of sequences of these symbols, subject to certain conditions. We will impose the following conditions: we assume that for each symbol i we allow certain symbols (depending only on i) to follow i and disallow all other symbols.

This information is best recorded in a $k \times k$ matrix A with entries in $\{0, 1\}$. That is, we allow the symbol j to follow the symbol i if and only if the corresponding (i, j)th entry of the matrix A (denoted by $A_{i,j}$) is equal to 1.

Definition. Let A be a $k \times k$ matrix with entries in $\{0, 1\}$. Let

$$\Sigma_A^+ = \{(x_j)_{j=0}^{\infty} \mid A_{x_i, x_{i+1}} = 1, \text{ for } j \in \mathbb{N}\}$$

denote the set of all infinite sequences of symbols (x_j) where symbol j can follow symbol i precisely when $A_{i,j} = 1$. We call Σ_A^+ a *(one-sided) shift of finite type*.

Let

$$\Sigma_A = \{(x_j)_{j=-\infty}^{\infty} \mid A_{x_j, x_{j+1}} = 1, \text{ for } j \in \mathbb{Z}\}$$

denote the set of all bi-infinite sequences of symbols subject to the same conditions. We call Σ_A a *(two-sided) shift of finite type*.

Sometimes for brevity we refer to Σ_A^+ or Σ_A as a *shift space*.

An alternative description of Σ_A^+ and Σ_A can be given as follows. Consider a graph with vertex set $\{1, 2, \ldots, k\}$ and with a directed edge from vertex i to vertex j precisely when $A_{i,j} = 1$. Then Σ_A^+ and Σ_A correspond to the set of all infinite (respectively, bi-infinite) paths in this graph.

It is possible for Σ_A^+ to be rather small. For example, if

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

then any occurrence of the symbol 1 must be followed by the symbol 2. However, no symbol can follow 2. Hence $\Sigma_A^+ = \emptyset$.

At the other extreme, if

$$A = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & \ddots & & \vdots \\ \vdots & & \ddots & \\ 1 & \cdots & & 1 \end{pmatrix},$$

the matrix where all entries are equal to 1, then there are no restrictions on which symbol can follow which other symbol. In this case we obtain the full one-sided and full two-sided k-shifts, Σ_k^+, Σ_k, respectively.

The following conditions on A guarantee that Σ_A^+ (or Σ_A) is large.

Definition. Let A be a $k \times k$ matrix with entries in $\{0, 1\}$. We say that A is *irreducible* if for each $i, j \in \{1, 2, \ldots, k\}$ there exists $n = n(i, j) > 0$ such that $(A^n)_{i,j} > 0$. (Here, $(A^n)_{i,j}$ denotes the (i, j)th entry of the nth power of A.)

Definition. Let A be a $k \times k$ matrix with entries in $\{0, 1\}$. We say that A is *aperiodic* if there exists $n > 0$ such that for all $i, j \in \{1, 2, \ldots, k\}$ we have $(A^n)_{i,j} > 0$.

In graph-theoretic terms, the matrix A is irreducible if there exists a path along edges from any vertex to any other vertex. The matrix A is aperiodic if this path can be chosen to have the same length (i.e. consist of the same number of edges), irrespective of the two vertices chosen.

We ill be interested in equipping shift spaces with a metric. What does it mean for two sequences in Σ_A^+ to be 'close'? Heuristically we will say that two sequences $(x_j)_{j=0}^\infty$ and $(y_j)_{j=0}^\infty$ are close if they agree for a large number of initial places.

More formally, for two sequences $x = (x_j)_{j=0}^\infty, y = (y_j)_{j=0}^\infty \in \Sigma_A^+$ we define

$$n(x, y) = \begin{cases} \sup\{n \mid x_j = y_j \text{ for } j = 0, 1, \ldots, n-1\} & \text{if } x \neq y \\ \infty & \text{if } x = y. \end{cases}$$

Thus $n(x,y)+1$ is the first place in which the sequences x and y disagree.

Fix $\theta \in (0,1)$. We define a metric d_θ on Σ_A^+ by

$$d_\theta((x_j)_{j=0}^\infty, (y_j)_{j=0}^\infty) = \theta^{n(x,y)}.$$

One can check that this is indeed a metric. The exact choice of $\theta \in (0,1)$ is, for the moment, unimportant. Each choice of θ gives the same topology on Σ_A^+.

In the two-sided case, we define for $x = (x_j)_{j=-\infty}^\infty, y = (y_j)_{j=-\infty}^\infty \in \Sigma_A$

$$n(x,y) = \begin{cases} \sup\{n \mid x_j = y_j \text{ for } |j| = 0,1,\ldots,n-1\} & \text{if } x \neq y \\ \infty & \text{if } x = y. \end{cases}$$

For a fixed choice of $\theta \in (0,1)$, we again define a metric d_θ on Σ_A in the same way:

$$d_\theta((x_j)_{j=-\infty}^\infty, (y_j)_{j=-\infty}^\infty) = \theta^{n(x,y)}.$$

Thus we can equip Σ_A and Σ_A^+ with a topology. It turns out that this topology is rather intricate.

Proposition 1.4
Suppose that A is an aperiodic matrix. Then, with the metrics defined as above, both Σ_A and Σ_A^+ are compact, totally disconnected, perfect metric spaces.

Remarks.

(i) A metric space X is *totally disconnected* if, for all $x \in X$, the connected component of X containing x is $\{x\}$.

(ii) A metric space X is *perfect* if it is equal to its limit points. Recall that a point $x \in X$ is a *limit point* if there exists a sequence of points x_n such that $x_n \to x$ but $x_n \neq x$.

(iii) The middle-third Cantor set is another example of a compact, totally disconnected, perfect metric space. Indeed, one defines a *Cantor set* to be any compact, totally disconnected, perfect metric space.

Define

$$\sigma^+ : \Sigma_A^+ \to \Sigma_A^+$$

by

$$(\sigma^+(x))_j = x_{j+1}.$$

Then σ^+ takes a sequence in Σ_A^+ and shifts it one place to the left (deleting the first term), We call σ^+ the *(one-sided, left) shift map*. (Often, if it is clear whether we are working with the one-sided or two-sided shift, we will drop the + in the superscript.)

There is a corresponding shift map on the two-sided shift space. Define

$$\sigma : \Sigma_A \to \Sigma_A$$

by

$$(\sigma(x))_j = x_{j+1},$$

so that σ shifts sequences one place to the left. Notice that in this case, we do not need to delete any terms in the sequence, We call σ the *(two-sided, left) shift map.*

Notice that σ is invertible but σ^+ is not.

When A is irreducible the shift map has strong dynamic properties.

Proposition 1.5
Suppose that A is aperiodic. Then both σ and σ^+ are chaotic: they have (i) sensitive dependence on initial conditions, (ii) the set of periodic points is dense, (iii) there is a dense orbit.

§1.7 Coding hyperbolic dynamical systems

As we have seen, the full one-sided 2-shift can be used to code the dynamics of the doubling map. This coding can then be used to prove dynamic properties of the doubling map, given that we know them for the shift. More generally, there is a class of dynamical systems—namely, those possessing some notion of 'hyperbolicity'—that can be modelled by an aperiodic shift of finite type. As the combinatorial definition of aperiodic shifts of finite type make them particularly tractable objects to study, we can then deduce many properties of hyperbolic dynamical systems. This will be the focus of later lectures.

Roughly speaking, a hyperbolic dynamical system on a space X is *hyperbolic* if through each point $x \in X$ there is sub-manifold $W^s(x)$ such that if $y \in W^s(x)$ then the orbits of x, y converge in forward time. If the dynamical system is invertible, then we will also require that through x there is another submanifold $W^u(x)$ transverse to $W^s(x)$ such that if $y \in W^u(x)$ then the orbits of x, y converge in backwards time.

For example, suppose T is that CAT map. This is given by a 2×2 matrix A with eigenvalues $\lambda_u = (3 + \sqrt{5})/2$ and $\lambda_s = (3 - \sqrt{5})/2$. Through each point $x \in \mathbb{R}^2/\mathbb{Z}^2$ there are two submanifolds: one, $W^s(x)$, parallel to λ_s, and one, $W^u(x)$, parallel to λ_u. One can easily check that these submanifolds have the above properties.

One can then show the existence of *Markov partitions*. A Markov partition is a partition of the phase space X constructed in such a way that if, for each $x \in X$, we record the sequence of partition elements that the orbit of x visits then we obtain an aperiodic shift of finite type. For the doubling map a Markov partition was easy to find: we used the partition

$X = [0, 1/2] \cup [1/2, 1]$ and then obtained the full one-sided 2-shift as a symbolic model. For more general hyperbolic dynamical systems it may not be obvious how to construct such a partition.

Bowen proved that if $T : X \to X$ is one of a particular wide class of hyperbolic dynamical system then there exists an aperiodic shift of finite type $\sigma : \Sigma_A \to \Sigma_A$ and a projection map $\pi : \Sigma \to X$ such that $T\pi = \pi\sigma$, and π is continuous, surjective and one-to-one except on a small set. (The precise notion of 'small' here need not concern us; the small set turns out to be the preimages of the boundary of the partition, as with the doubling map.) One can then use this symbolic coding to deduce man ergodic theoretic properties of hyperbolic dynamical systems.

§1.8 References

A good introduction to many different topics in dynamical systems can be found in

A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Encyclopædia of Math., vol. 54, C.U.P., Cambridge, 1995.

Good introductions to ergodic theory are

W. Parry, *Topics in Ergodic Theory*, C.U.P., Cambridge, 1981.

K. Petersen, *Ergodic Theory*, C.U.P., Cambridge, 1983.

P. Walters, *An introduction to ergodic theory*, Springer, Berlin, 1982.

The idea of coding a dynamical system symbolically goes back to at least Koebe in the 1920s and Hedlund in the 1930s, at least for the geodesic flow on surfaces of constant negative curvature. See

C. Series, *Geometrical Markov coding on surfaces of constant negative curvature*, Ergod. Th. & Dynam. Sys. **6** (1986), 601–625.

for historical comments. A thorough treatment in the context of the ergodic theory of hyperblic dynamical systems is given in

R. Bowen, *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, Lecture Notes in Math., vol. 470, Springer, Berlin, 1975.

Connections between ergodic theory and hyperbolic geometry via hyperbolic dynamics are described in

T. Bedford, M. Keane, and C. Series (eds.), *Ergodic theory, symbolic dynamics and hyperbolic spaces*, O.U.P., Oxford, 1991.

§1.9 Exercises

Exercise 1.1
Let $T(x) = x + \alpha$ be an irrational rotation of the circle. Prove that every orbit of T is dense in $\mathbb{R}/\mathbb{Z}$.

Exercise 1.2
Let A be a $k \times k$ $0-1$ matrix. Let $\theta \in (0,1)$.

(i) Check that d_θ is a metric on Σ_A^+.

(ii) Show that Σ_A^+ is a closed subset of the compact set Σ_k.

(iii) Show that Σ_A^+ is totally disconnected.

(iv) Check that σ is continuous.

One can extend (ii) and (iii) further and show that if A is irreducible then Σ_A^+ is a compact, perfect (that is, is equal to its limit points) totally disconnected set. As such, topologically it is a Cantor set.

Exercise 1.3
(i) Show that the full one-sided 2-shift Σ_2^+ is chaotic in the sense of Devaney. Hence conclude that the doubling map is chaotic.

(ii) Conclude the same for the transformation $x \mapsto \beta x \bmod 1$ where $\beta > 1$ is the golden mean ($\beta^2 = \beta + 1$), using the shift of finite type Σ_A^+ with matrix

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