Hyperbolic sets and Markov partitions

Consider a map f defined in R^n. Assume that in R^n there exist m mutually non-intersecting "rectangles" $\Pi_1, \ldots, \Pi_m$ such that $\Pi_i = X_i \times Y_i$, where X_i are closed bounded convex subsets of R^{m_1} and Y_i are closed bounded convex subsets of R^{m_2} ($i = 1, \ldots, m$ and $m_1 + m_2 = m$), and that for each pair of i and j such that $i, j = 1, \ldots, m$ there exist N_{ij} pairs of smooth functions $p_{s,ij}, q_{s,ij}$ ($s = 1, \ldots, N_{ij}$, if $N_{ij} = 0$ then the functions p, q are not defined for these i, j) such that a point $(x, y) \in \Pi_i$ is mapped by f to a point $(\bar{x}, \bar{y}) \in \Pi_j$ if and only if $N_{ij} \neq 0$ and

$$\bar{x} = p_{s,ij}(x, \bar{y}), \qquad y = q_{s,ij}(x, \bar{y}) \tag{1}$$

for some $s = 1, \ldots, N_{ij}$. This relation between $(x, \bar{y})$ and $(\bar{x}, y)$ is the so-called *cross-form* for the map f, and it is very convenient for the analysis of the dynamics near hyperbolic sets.

Let the following 3 properties hold:
A) for all i, j, s

$$p_{s,ij}(X_i, Y_j) \subseteq X_j, \qquad q_{s,ij}(X_i, Y_j) \subseteq Y_i; \tag{2}$$

B) for any i, j equation

$$\begin{cases} p_{s_1,ij}(u, v) = p_{s_2,ij}(u, v) \\ q_{s_1,ij}(u, v) = q_{s_2,ij}(u, v) \end{cases}$$

has no solutions (u, v) if $s_1 \neq s_2$;
C) there exists $\lambda < 1$ such that for all i, j, s

$$\left\| \frac{\partial p}{\partial x} \right\|_\circ + \left\| \frac{\partial p}{\partial \bar{y}} \right\|_\circ \leq \lambda, \qquad \left\| \frac{\partial q}{\partial x} \right\|_\circ + \left\| \frac{\partial q}{\partial \bar{y}} \right\|_\circ \leq \lambda, \tag{3}$$

where $\| \cdot \|_\circ$ denotes the supremum of the norm over all $x \in X_i$, $\bar{y} \in Y_j$.

It follows from properties A and B that the intersection $\Pi_i \cap f^{-1}(\Pi_j)$ consists of N_{ij} connected components:

$$\Pi_i \cap f^{-1}(\Pi_j) = \Pi_{1,ij} \cup \ldots \cup \Pi_{N_{ij},ij}.$$

Indeed, by (1)and (2), the intersection $\Pi_i \cap f^{-1}(\Pi_j)$ is foliated by the preimages of the lines (or planes, if $m_1 > 1$) $y = c = const \in Y_j$, i.e. the surfaces

$$y = q_{s,ij}(x, c). \tag{4}$$

Fix any $s = 1, \ldots, N_i j$ and denote as $\Pi_{s,ij}$ the union of all corresponding lines (i.e. c runs all Y_j for the given value of s). Condition B guarantees that $\Pi_{s_1,ij}$ and $\Pi_{s_2,ij}$ do not intersect if $s_1 \neq s_2$: would (x, y) be an intersection point, that would mean $(x, y) \in \Pi_i$ and there would exist $(\bar{x}, \bar{y}) = f(x, y) \in \Pi_j$ such that

$$x = p_{s_1,ij}(x, \bar{y}) = p_{s_2,ij}(x, \bar{y}), \qquad y = q_{s_1,ij}(x, \bar{y}) = q_{s_2,ij}(x, \bar{y}),$$

a contradiction.

Since $\|q'_x\| \leq \lambda$ by (3), every surface (or line, if $m_1 = 1$) given by (4) is Lipshitz with the Lipshitz constant λ or less. We call such surfaces/lines horizontal, and the sets $\Pi_{s,ij}$ are, therefore, called horizontal strips. As $\|p'_x\| < 1$, it follows that the map f, when restricted to any of these surfaces (recall they correspond to $\bar{y} = c$), is contracting. Similarly to the above, the intersection $f(\Pi_i) \cap \Pi_j$ consists of N_{ij} vertical strips, foliated by the vertical surfaces/lines of the form

$$x = p_{s,ij}(y, c), \quad y \in Y_j,$$

where c may take arbitrary values in X_i, and the map f^{-1} is contracting in restriction to any of these surfaces.

As we see, condition C implies that the map f is contracting in some directions (x-directions) and expanding in complementary directions (y-directions, since f^{-1} is contracting in y, it means that f is expanding in y). This property is called *hyperbolicity*, so condition 3 is a way of formulating the hyperbolicity property.

Let Λ be a set of all points whose entire orbits stay in $\Pi_1 \cup \ldots \Pi_m$, i.e. $M \in \Lambda$ if and only if for every integer k, positive or negative, there exists i_k such that $M_k = f^k M \in \Pi_1 \cup \ldots \Pi_m$. Note that $M_k \in f^{-1}(\Pi_{i_{k+1}}) \cap \Pi_{i_k}$, hence for every k we have $N_{i_k i_{k+1}} > 0$ and there exists a uniquely defined s_k such that $M_k \in \Pi_{s_k, i_k i_{k+1}}$. We will call the sequence of symbols $\{\alpha_k = (i_k, s_k)\}_{k=-\infty}^{+\infty}$ the code $\alpha = \psi(M)$ of the point M.

Let us build an oriented graph G with m vertices such that there is exactly N_{ij} edges which from vertex i to vertex j. Since $s_k = 1, \ldots, N_{i_k i_{k+1}}$, it follows that the code α can be viewed as an infinite path along the edges of G. Thus, when conditions A and B are satisfied, we define the coding map ψ from Λ to the set of infinite paths along the edges of the graph G. We assume that the map ψ is continuous with continuous inverse. Then the coding map is continuous as well: if two points are sufficiently close, then their orbits stay close for sufficiently many iterations both forward and backward (by the continuity of f and f^{-1}), hence they visit the same strips $\Pi_{s_k, i_k i_{k+1}}$ for all $|k| \leq K$ where K is sufficiently large; in other words, if two points in Λ are close, their codes coincide for a sufficiently long segment of k values, which means the continuity of the coding map.

Note that the following identity is, by construction, satisfied for every point in Λ:

$$\psi \circ f = S \circ \psi, \tag{5}$$

where S stands for the shift map on the space of infinite paths in G, namely $(S\alpha)_k = \alpha_{k+1}$. In words, this relation reads as "f on Λ is semi-conjugate to S". As we showed, this semi-conjugacy (i.e. the map ψ) is continuous.

When the inverse to the map ψ exists, this map is called "conjugacy" rather than semi-conjugacy, and we say that "f is conjugate to S", or we say "f is topologically conjugate to S" to stress that ψ is a homeomorphism (i.e. both ψ and ψ^{-1} are continuous; if ψ is a continuous map of a Hausdorff compact, then the existence of the inverse map also implies the continuity of the inverse).

Note that if f is conjugate to S, then S is also conjugate to f (since (5) implies $f \circ \psi^{-1} = \psi^{-1} \circ S$ which means that ψ^{-1} is a conjugacy between S and f). Note also that (5) implies $\psi \circ f^k = S^k \circ \psi$, i.e the conjugacy takes orbits of f to orbits of S.

Topological conjugacy is a basic notion of the dynamical systems theory: if two systems are topologically conjugate, we believe that it is safe to say they have the same dynamics. In particular, if two maps are conjugate, periodic points of one map correspond to periodic points of the other, and if an orbit of one map tends, say, to a periodic orbit, then the image, by the conjugacy, of the first orbit tends to the image of the limit periodic orbit, etc..

In fact, when the hyperbolicity condition C is satisfied, the semi-conjugacy between our map f and the shift map S is a topological conjugacy indeed, as the following theorem shows.

Theorem 1. *The coding map ψ is a homeomorphism. Namely, for every infinite path along the edges of graph G there is a unique point in Λ which has this path as a code, and if two paths close (i.e. if they coincide on a sufficiently long segment around the starting point), then the points in Λ that correspond to them are close as well.*

Proof. A point $M(x_0, y_0)$ with the orbit $f^k M = M_k(x_k, y_k)$, $k = -\infty, \dots, +\infty$, has a code $\alpha = \{(s_k, i_k)\}_{k=-\infty}^{+\infty}$ if and only if for each k

$$x_{k+1} = p_{s_k, i_k i_{k+1}}(x_k, y_{k+1}), \qquad y_k = q_{s_k, i_k i_{k+1}}(x_k, y_{k+1}). \tag{6}$$

Thus, we will prove that each code corresponds to a certain, uniquely defined point in Λ if we show that the system of equations (6) (taken together for all k) has a unique solution for every choice of the sequences i_k and s_k ($1 \le s_k \le N_{i_k i_{k+1}}$).

In order to do so, consider the operator Φ_α which takes a sequence (x_k, y_k), $k = -\infty, \dots, +\infty$, to the sequence $(\hat{x}_k, \hat{y}_k)$ defined by the following rule:

$$\hat{x}_{k+1} = p_{s_k, i_k i_{k+1}}(x_k, y_{k+1}), \qquad \hat{y}_k = q_{s_k, i_k i_{k+1}}(x_k, y_{k+1}). \tag{7}$$

The operator Φ_α acts on the set D_α of sequences satisfying $x_k \in X_{i_k}$, $y_k \in Y_{i_k}$; it immediately follows from condition A that $\Phi_\alpha(D_\alpha) \subseteq D_\alpha$.

Introduce the following metric on D_α:

$$dist(\{(x_k^{(1)}, y_k^{(1)})\}, \{(x_k^{(2)}, y_k^{(2)})\}) = \sup_k \max\{\|x_k^{(1)} - x_k^{(2)}\|, \|y_k^{(1)} - y_k^{(2)}\|\}. \tag{8}$$

Since the sets X_{i_k}, Y_{i_k} are closed subsets of complete spaces (R^{m_1} and R^{m_2}), they are complete as well, and this implies that A_α is complete, when endowed with metric (8). Thus, we will have the existence and uniqueness of the fixed point of the operator Φ_α (this fixed point is the sought solution of system (6), obviously) if we show that Φ_α is contracting in metric (8).

It remains to note that the contraction property easily follows from condition C. Indeed, let us take any two sequences $\{(x_k^{(1)}, y_k^{(1)})\}$ and $\{(x_k^{(2)}, y_k^{(2)})\}$ from

D_α and let d be the distance between them. Then, by applying the Mean Value Theorem to (7), we find

$$\|\hat{x}^{(1)}_{k+1} - \hat{x}^{(2)}_{k+1}\| \leq \left\|\frac{\partial p_{s_k, i_k i_{k+1}}}{\partial x}\right\|_\circ \|\hat{x}^{(1)}_k - \hat{x}^{(2)}_k\| + \left\|\frac{\partial p_{s_k, i_k i_{k+1}}}{\partial \bar{y}}\right\|_\circ \|\hat{y}^{(1)}_{k+1} - \hat{y}^{(2)}_{k+1}\|$$

(we use here that the sets X_i, Y_i are convex). By (3),(8), we now obtain

$$\|\hat{x}^{(1)}_{k+1} - \hat{x}^{(2)}_{k+1}\| \leq \lambda\, d.$$

Similarly,

$$\|\hat{y}^{(1)}_k - \hat{y}^{(2)}_k\| \leq \lambda\, d,$$

and we find that

$$dist(\{(\hat{x}^{(1)}_k, \hat{y}^{(1)}_k)\}, \{(\hat{x}^{(2)}_k, \hat{y}^{(2)}_k)\}) \leq \lambda\, dist(\{(x^{(1)}_k, y^{(1)}_k)\}, \{(x^{(2)}_k, y^{(2)}_k)\}),$$

which means contraction as $\lambda < 1$.

Thus, given any infinite path α along the edges of the graph G, we indeed have (by the Banach contraction mapping principle) the existence and uniqueness of a point $M \in \Lambda$ with the code α. To complete the theorem, we need to prove that M depends on α continuously. To this aim, we prove the following statement: there exists a constant C such that for any points $M^{(1)}$ and $M^{(2)}$ in Λ if their codes $\alpha^{(1)} = \psi(M^{(1)})$ and $\alpha^{(2)} = \psi(M^{(2)})$ coincide on the centered at zero segment of length $2K+1$:

$$\alpha^{(1)}_k = \alpha^{(2)}_k \qquad \text{for all} \quad |k| \leq K,$$

then

$$dist(M^{(1)}, M^{(2)}) \leq C\lambda^K, \tag{9}$$

where we denote $dist((x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)})) = \max\{\|x^{(1)} - x^{(2)}\|, \|y^{(1)} - y^{(2)}\|\}$.

We prove the claim by induction in K. It is trivially true for $K = 0$ — just take C bigger then the distance between any two points in $\Pi_1 \cup \ldots \cup \Pi_m$. Thus, if we denote $d_1 = dist(f(M^{(1)}), f(M^{(2)}))$, then we may assume

$$d_1 \leq C\lambda^{K-1}, \tag{10}$$

since the codes for the points $f(M^{(1)})$ and $f(M^{(2)})$ coincide on the centered at zero segment of length at lest $2K-1$ (these codes are obtained by the shift S applied to the codes of $M^{(1)}$ and $M^{(2)}$). Similarly, we may assume

$$d_{-1} \leq C\lambda^{K-1}, \tag{11}$$

where $d_{-1} = dist(f^{-1}(M^{(1)}), f^{-1}(M^{(2)}))$. By induction, we will have (9) proven for all K, if we prove that for every given K estimates (10) and (11) imply (9).

Let $M^{(1)} = (x^1_0, y^{(1)}_0)$, $M^{(2)} = (x^2_0, y^{(2)}_0)$, $f(M^{(1)}) = (x^1_1, y^{(1)}_1)$, $f(M^{(2)}) = (x^2_1, y^{(2)}_1)$, $f^{-1}(M^{(1)}) = (x^1_{-1}, y^{(1)}_{-1})$, $f^{-1}(M^{(2)}) = (x^2_{-1}, y^{(2)}_{-1})$. By (1),

$$x_0 = p(x_{-1}, y_0), \qquad y_0 = q(x_0, y_1),$$

therefore, by (3),

$$\|x_0^1 - x_0^2\| \leq \lambda \max\{\|x_{-1}^1 - x_{-1}^2\|, \|y_0^{(1)} - y_0^{(2)}\|\} \leq \lambda \max\{d_{-1}, d_0\},$$

$$\|y_0^1 - y_0^2\| \leq \lambda \max\{\|x_0^1 - x_0^2\|, \|y_1^{(1)} - y_1^{(2)}\|\} \leq \lambda \max\{d_0, d_1\},$$

where we denote $d_0 = dist(M^{(1)}, M^{(2)})$. It follows that

$$d_0 \leq \lambda \max\{d_0, d_{-1}, d_1\},$$

which implies

$$d_0 \leq \lambda \max\{d_{-1}, d_1\},$$

hence, estimates (10) and (11) imply (9) indeed.

We see that if the length of the centered at zero segment on which two codes coincide tends to infinity, then the distance between the corresponding points in Λ tends to zero, which means the continuity of ψ^{-1}, as required. $\square$

As we explained before the theorem, this result means that the dynamics of the map f on the set Λ is equivalent to that of the dynamics of the shift map S, i.e. to the walk along the edges of the graph G. Note that once we arrive to a vertex i of the graph, we can move to any vertex j which is connected with i by an edge of G, and the choice of where to move is not restricted by where we were before we came to the vertex i, i.e. the further motion depends on the current position only and is independent of the history. This is called a "Markov property", in analogy to a similar property considered in the probability theory. Therefore the set of the rectangles $\Pi_1, \ldots, \Pi_m$ which satisfy conditions A,B and C is called a Markov partition for the set *Lambda*.

The set Λ is called a hyperbolic set. By construction, Λ is invariant with respect to f (recall that Λ is the set of all points whose entire orbits never leave $\Pi_1 \cup \ldots \cup \Pi_m$, so, by the definition, if $M \in \Lambda$, then its whole orbit is in Λ as well). The set λ is also bounded (since $\Lambda \subset \Pi_1 \cup \ldots \cup \Pi_m$ and the rectangles Π_i are bounded) and closed (since the property of not being in Λ is open: if a point gets outside $\Pi_1 \cup \ldots \cup \Pi_m$, then all close points get outside as well, because of the closeness of Π_i's). Thus, Λ is a compact invariant set.

It also follows from Theorem 1 that Λ is zero-dimensional: it is homeomorphic to the set of all paths along the edges of the graph G, and the latter set is zero-dimensional, obviously. Recall that a set is zero-dimensional, if for any point in it there exist arbitrarily small neighbourhoods which are disconnected from the rest of the set. For a subset of R^m this can be reformulated as follows: the set is zero-dimensional, if for any point M in it there exist arbitrarily small neighbourhoods of M whose boundary does not belong to the set. A way to check the zero-dimensionality is, for any point M and any $\varepsilon > 0$, to find a function ξ, which is continuous on the set, takes only two values: 0 and 1, equals to zero outside the ε-neighbourhood of M and equals to 1 at M. Let us show how such functions can be found for the set Λ. Take any $M \in \Lambda$ and any $\varepsilon > 0$. Let K be such that $C\lambda^K < \varepsilon$, i.e. every point whose code coincides with the code of M on the centered at zero segment of length $2K+1$ is ε-close to M (see (9)). Define $\xi(M')=1$ if the code of the point $M'\in\Lambda$ coincides with the code of M on the centered at zero segment of length $2K+1$, and let $\xi(M')=0$. Obviously, the function ξ satisfies all the above requirements (it is continuous on Λ, since it depends continuously on the code, and the code is a continuous function of the point in Λ).

As we see, the Markov partition in the way we introduced it provides a tool for constructing zero-dimensional compact invariant hyperbolic sets. Note that there are other ways of defining hyperbolic sets. However one may show that our construction of zero-dimensional hyperbolic sets is equivalent to the others.

Example. Consider the Henon map $(x,y)\mapsto(\bar{x},\bar{y})$:

$$\left\{\begin{array}{l}\bar{x}=y,\\ \bar{y}=M-x-y^2,\end{array}\right. \tag{12}$$

where M is a parameter. Choose some $a>0$ and consider a rectangle Π_1 : $\{|x|\le a, |y|\le a\}$ (i.e. we take $X_1=Y_1=[-a,a]$). Rewriting map (12) in the cross-form (1), we obtain

$$\left\{\begin{array}{l}\bar{x}=y,\\ y=\pm\sqrt{M-x-\bar{y}},\end{array}\right. \tag{13}$$

i.e. we have here 2 pairs of functions, $p_{1,11}, q_{1,11}$ and $p_{2,11}, q_{2,11}$, where

$$p_{1,11}=q_{1,11}=+\sqrt{M-x-\bar{y}}, \qquad p_{2,11}=q_{2,11}=-\sqrt{M-x-\bar{y}}.$$

Let us find out when conditions A,B,C are satisfied by these functions.

To satisfy condition A, the functions p and q must be defined for all $|x|\le a$, $|\bar{y}|\le a$, and their values must not exceed a in the absolute value. This gives us conditions

$$M-2a\ge 0$$

and

$$\sqrt{M+2a}\le a$$

. Condition B reads as the absence of the solutions to

$$+\sqrt{M-u-v}=-\sqrt{M-u-v}$$

at $|u|\le a$, $|v|\le a$, i.e. we must require

$$M-2a>0.$$

Condition C reads as

$$\sup_{|x|\le a,|\bar{y}|\le a}\frac{1}{\sqrt{M-x-\bar{y}}}<1,$$

hence

$$M-2a>1.$$

As we see, Π_1 is a Markov partition if

$$2a + 1 < M \le a^2 - 2a. \tag{14}$$

This gives us that at $M > 5 + 2\sqrt{5}$ the Henon map (12) has a hyperbolic invariant set Λ (the set of all points whose orbits never leave the rectangle Π_1 with $a = 1 + \sqrt{M+1}$). Moreover, the map on Λ is topologically conjugate to the shift on the set of all paths along the graph G with 1 vertex and two edges (loops, which start at and return to the same vertex). Thus, the orbits in Λ are in one-to-one correspondence with all infinite sequences of two symbols (one symbol corresponds to a passage along one of the loops in the graph, and the other symbol corresponds to a passage along the second loop). Such hyperbolic set is called a Smale horseshoe. It is easy to check that when condition (14) is fulfilled, the image of Π_1 indeed has a horseshoe-like shape, and intersects Π_1 at two vertical strips (the preimage of Π_1 also has a horseshoe-like shape, and its intersection with Π_1 consists of two horizontal strips).

Transitivity, periodic points, and topological entropy

An orientable graph G is called transitive if for any two of its vertices there is a path which connects one vertex with the other. We further consider only such (finite) graphs for which at least one edge emanates from every vertex of the graph, and at least one edge enters each vertex. Obviously, any such graph can be decomposed into transitive components: subgraphs $G_1, \dots, G_\ell$ such that each vertex of G belongs to one of these subgraphs, each of the subgraphs is transitive, and if in G there is a path which leads from G_{r_1} to G_{r_2}, then in G there is no path which leads from G_{r_2} to G_{r_1}.

Let us have a Markov partition $\Pi_1, \dots, \Pi_m$ for a zero-dimensional hyperbolic set Λ. Consider the graph G defined by the partition. Let $G_1, \dots, G_\ell$ be the decomposition of G into the transitive components. Let n_r be the number of vertices in the component G_r (so $n_1 + \ldots + n_\ell = m$). We may always enumerate the rectangles in the partition in such a way that the vertices $1, \dots, n_1$ will belong to G_1, the vertices $n_1 + 1, \dots, n_1 + n_2$ will belong to G_2, etc.. Denote $m_0 = 0$, $m_s = m_{r-1} + n_r$, $r = 1, \dots, \ell$, so the vertices $m_{r-1} + 1, \dots, m_r$ are exactly those which belong to the component G_r.

Now, denote as Λ_r the set of all points whose orbits never leave $\tilde{\Pi}_r = \Pi_{m_{r-1}+1} \cup \ldots \cup \Pi_{m_r}$. Each of the sets Λ_r is a compact invariant hyperbolic set, and consists exactly of those points in Λ whose codes are the paths along the edges of the subgraph G_r. The orbits in Λ which do not belong to $\Lambda_1 \cup \ldots \cup \Lambda_\ell$ correspond to the paths which go from one transitive component of G to another and never come back. Therefore, these orbits are "wandering": they leave a group $\tilde{\Pi}_{r_1}$ of the rectangles and enter another group, $\tilde{\Pi}_{r_2}$, and then never return to $\tilde{\Pi}_{r_2}$. Since the number of the transitive components is finite, any such orbit eventually enters one of the rectangle's groups $\tilde{\Pi}_r$ and stays there forever. In fact, this orbit tends (exponentially) to some orbit from the set Λ_r: if we denote the orbit under consideration as $\{M_k = f^k(M_0)\}_{k=-\infty}^{+\infty}$, then in Λ_r there exists a point M' such that

$$dist(M_k, f^k(M')) = O(\lambda^k).$$

Indeed, we have that after a certain moment, the code α of the point M_0 contains only the edges of the subgraph G_r: the code $\alpha = \{\alpha_k = (s_k, i_k)\}_{k=-\infty}^{+\infty}$ satisfies $m_{r-1} + 1 \le i_k \le m_r$ for all $k \ge k_0$ (which means that these vertices i_k belong to G_r). Now, take a finite path γ in G_r which connects the vertex i_{k_0} with itself (such exists by the transitivity of G_r), and build an infinite path β by concatenation of a periodically repeated path γ with the right tail $\alpha_{k_0}\alpha_{k_0+1}\ldots$ of α, i.e. $\beta = \ldots\gamma\gamma\gamma\alpha_{k_0}\alpha_{k_0+1}\alpha_{k_0+2}\ldots$. By construction, the path β is contained within G_r, hence the point M' with the code β (such exists by Theorem 1: $M' = \psi^{-1}(\beta)$) lies in Λ_r. Since the codes (α and β) of the points M_0 and M' have equal tails, the exponential convergence of $f^k(M_0)$ to $f^k(M')$ follows from estimate (9).

Thus every orbit from Λ which does not belong to one of the sets Λ_r, $r = 1, \dots, \ell$, tends to one of the orbits of one of these sets. This means that the analysis of the dynamics in the set Λ is reduced to the analysis of the dynamics in the subsets Λ_r. These subsets are called the transitive components of Λ. The rectangles $\Pi_{m_{r-1}+1}, \dots, \Pi_{m_r}$ form a Markov partition for the set Λ_r, and the corresponding graph is G_r. In general, when the graph which is defined by a Markov partition is transitive, the corresponding hyperbolic set is called transitive.

Theorem 2. *Let Λ be a transitive zero-dimensional hyperbolic set. Then there exists $M \in \Lambda$ such that the orbit of M is dense in Λ. Periodic points are dense in Λ. If Λ consists of more than one orbit, then the set of orbits which are dense in Λ is uncountable, and the set of periodic orbits is infinite (countable); moreover, in this case, given any two periodic orbits in Λ, the set of points of heteroclinic orbits which connect them is infinite (countable) and dense in Λ.*

Proof. The orbit $f^k(M)$ is dense in a set if it visits any neighbourhood of any point in this set. By Theorem 1, the point $f^k(M)$ lies, for some k, in a small neighbourhood of a point $M' \in \Lambda$, if and only if the codes of these points, $\psi(f^k(M))$ and $\psi(M')$, coincide on some sufficiently long segment centered at zero. Since the code of $f^k(M)$ is obtained from the code of M just by the shift S^k, it follows that the orbit of the point M is dense in Λ if and only if its code (an infinite path along the edges of the graph defined by the Markov partition) contains all finite paths along the edges of the graph.

So, we prove the first claim of the theorem by showing that such infinite path exists. Intuitively, the existence of such path follows immediately from the transitivity of the graph: by the transitivity condition, while walking along the edges we may at any instance move to any vertex we want and then continue along any path that emanates from this vertex for an arbitrary long time. To make this argument more formal, for each pair of vertices i, j denote as γ_{ij} a finite path that connects them (at least one such path exists by the transitivity). Now, take the set of all possible finite paths in the graph. This set is countable, i.e. we may enumerate the elements of this set. This would give us a sequence of finite paths $\dots, \beta_{-2}, \beta_{-1}, \beta_0, \beta_1, \dots$ which consists of all possible finite paths in the graph under consideration. Let i_t be the first vertex of the path β_t and j_t be its last vertex. Then the infinite path α obtained by the concatenation of consecutive paths $(\beta_t \gamma_{j_t i_{t+1}})$ is the sought path that contains all possible finite paths.

Similarly, as periodic points of f in Λ correspond to periodic paths in the graph (these are obtained by periodically repeating the same closed finite path), we prove that periodic points of f are dense in Λ by showing that the set of all periodic paths is a dense subset of the set of all infinite paths in the transitive graph. This amounts to showing that given any finite path β in a transitive graph, one can find a periodic path which includes β. The latter is trivial: if β starts at vertex i and ends at vertex j, then the path $\beta\gamma_{ji}$ is closed (it starts and ends at the same vertex i), so by repeating this closed path infinitely many times we obtain the sought periodic path that contains β.

Note that different finite paths β correspond to different periodic paths. A periodic orbit consists of a finite number of points and cannot alone be dense in the set Λ, unless Λ consists of this one orbit. However, as we showed, the set of all periodic points is dense in λ.

Now, let us consider the case where the transitive set Λ consists of more than one orbit. Since the orbits in Λ are in one-to-one correspondence to the infinite paths in the graph, it follows that there is more than one (up to a shift) infinite path in the graph. This implies that there is a vertex i_0 in the graph from which at least two edges emanate, s_1 and s_2. Let s_1 go to the vertex j_1 and s_2 go to the vertex j_2. Now take any vertices i and j and denote $\gamma_{ij}^1 = \gamma_{ii_0} s_1 \gamma_{j_1 j}$, $\gamma_{ij}^2 = \gamma_{ii_0} s_2 \gamma_{j_2 j}$. By construction, γ_{ij}^1 and γ_{ij}^2 are two different finite paths which connect i with j (the paths are different since $s_1 \neq s_2$).

Let, as above, $\ldots, \beta_{-2}, \beta_{-1}, \beta_0, \beta_1, \ldots$ be a sequence of finite paths in the graph under consideration which consists of all possible finite paths in the graph. Take any sequence $\sigma = \{\sigma_t\}_{t=-\infty}^{+\infty}$ of 1's and 2's. The path α_σ obtained by the concatenation of consecutive paths $(\beta_t \gamma_{j_t i_{t+1}}^{\sigma_t})$ (where i_t is the vertex where the path β_t starts and j_t is the vertex where it ends) contains all possible finite paths, hence the orbit of the point $M_\sigma \in \Lambda$ with the code α_σ (i.e. $M_\sigma = \psi^{-1}(\alpha_\sigma)$) is dense in Λ. By construction, the codes α_σ which correspond to different sequences σ do not coincide, therefore corresponding points M_σ are different, and the cardinality of the set of these points if the same as the cardinality of the set of all infinite binary sequences, i.e. it is uncountable, as was to be proved.

It follows that the set Λ is uncountable itself in this case. Since periodic points are dense in Λ, their number must be infinite (it is countable since the number of periodic points equals to the number of closed finite paths in the graph, i.e. it does not exceed the number of all finite paths in the graph, which is countable).

Now take any two periodic points $M^1 \in \Lambda$, $M^2 \in \Lambda$. A point M is called heteroclinic, if

$$\lim_{k \to -\infty} dist(f^k M, f^k M^1) = 0, \qquad \lim_{k \to +\infty} dist(f^k M, f^k M^2) = 0,$$

i.e. the orbit of M "connects" the orbit of M^1 with the orbit of M^2 (if M^1 and M^2 are points of the same orbit, then M is also called homoclinic). If the code of the periodic point M^1 is obtained by the infinite repetition of a finite closed path α^1 and the code of M^2 is obtained by the infinite repetition of a finite closed path α^2, then M is heteroclinic if and only if the right tail of its code coincides with the infinitely repeated α^2 and the left tail coincides with the infinitely repeated α^1. Thus, in order to prove that heteroclinics to the given periodic points M^1 and M^2 in the transitive hyperbolic set Λ form a dense subset of Λ, we must prove that given any two closed paths α^1 and α^2 in a transitive graph and any finite path β in the same graph we can always build an infinite path in this graph, which would contain β and its right tail would coincide with the infinitely repeated α^2 while the left tail would coincide with the infinitely repeated α^1. It remains to note that if α^1 ends at the vertex i_1, α^2 starts with the vertex i_2, β starts with i and ends at j, then the sought heteroclinic path is $\ldots\alpha^1\alpha^1\gamma_{i_1 i}\beta\gamma_{ji_2}\alpha^2\alpha^2\ldots$. $\square$

In general, an invariant set is called transitive, if it is a closure of one orbit. As we see, in the case of zero-dimensional hyperbolic sets the transitivity is equivalent to the transitivity of the corresponding graph, and implies that periodic, heteroclinic, homoclinic orbits are dense in the transitive set as well. Note also that the dynamics of the hyperbolic sets defined by their Markov partitions is very rich: unless we have a trivial case where each transitive component consists of a single periodic orbit, the number of different periodic orbits, for example, grows without bound as the period increases.

In order to characterise exactly how the number of periodic orbits grows with period, it is useful to introduce a transition matrix of the graph G, namely the matrix $\mathcal{N}$ with nonnegative entries N_{ij}, where N_{ij} is the number of edges of the graph which go from the vertex i to the vertex j. Obviously, the number of length-k paths which start with i and end at j is the entry i,j of the matrix $\mathcal{N}^k$. In particular, the number of closed paths of length k which start (and end) at the vertex i equals to $\mathcal{N}_{ii}^k$. By summing over all vertices, we find that the number of closed paths of length k equals to

$$\sum_{i=1,\ldots,m} \mathcal{N}_{ii}^k = tr\mathcal{N}^k = \sum_{i=1,\ldots,m} \lambda_i^k,$$

where $\lambda_1,\ldots,\lambda_m$ are the eigenvalues of the transition matrix $\mathcal{N}$. By Theorem 1, periodic points of the map f on Λ correspond to closed paths in the graph G, so we find

$$P_k = \sum_{i=1,\ldots,m} \lambda_i^k \tag{15}$$

where P_k is the number of periodic points of period k (note that we do not claim here that k is the least period of the point, e.g. fixed points are counted as periodic points of any period, points of period 2 are counted as periodic points of any even period, etc.).

Denote $\bar{\lambda} = \max\{|\lambda_1|,\ldots,|\lambda_m|\}$. One may infer from (15) that

$$\limsup_{k\to+\infty} \frac{\ln P_k}{k} = \ln\bar{\lambda} \tag{16}$$

(we omit the proof). Note that the lower limit may be smaller: for example, the matrix $\mathcal{N} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$ has eigenvalues $\sqrt{2}, -\sqrt{2}, 0, 0$, so $tr\mathcal{N}^k = \sqrt{2^k}(1+(-1)^k)$, i.e. P_k is zero for all odd k.

It follows from (16) that the number of periodic orbits grows exponentially with period, provided $\bar{\lambda} > 1$. One may show that the number of closed paths in any graph grows exponentially with period always, except for the case where every transitive component consists of a single periodic orbit (hence only in this case we have $\bar{\lambda} = 1$).

Denote as Q_k the number of all possible paths of length k in the graph G. Let us show that

$$\limsup_{k\to+\infty} \frac{\ln Q_k}{k} \le \ln \bar\lambda. \tag{17}$$

Indeed, Q_k is the sum of all the elements of the matrix $\mathcal{N}^k$, i.e.

$$Q_k = \|\mathcal{N}^k \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \|$$

where we define the norm $\|\cdot\|$ of a vector here as the sum of the absolute values of all its components. Thus,

$$Q_k \le \|\mathcal{N}^k\| m = m\|T\circ J^k \circ T^{-1}\| \le m\|T\|\ \|T^{-1}\|\ \|J^k\|$$

where J is the Jordan form of the matrix $\mathcal{N}$ and T defines the transformation to the Jordan form: $\mathcal{N} = T\circ J \circ T^{-1}$. It is easy to check that $\|J^k\| \le m\bar\lambda^k k^m$. Hence

$$Q_k = O(k^m \bar\lambda^k),$$

from which (**??**) follows. Since $Q_k \ge P_k$, we extract from (16),(17) that

$$h = \limsup_{k\to+\infty} \frac{\ln Q_k}{k} = \ln \bar\lambda. \tag{18}$$

The number h is called the topological entropy (of the shift map on the set of infinite paths in the graph G). In general, given a map f and an invariant set Λ, the topological entropy of $f|_\Lambda$ is defined as follows. For every $\varepsilon > 0$, $n > 0$, let $K(\varepsilon, n)$ be the maximal possible number of ε-separated orbits of length n in Λ (the orbits $\{f^k(M^1)\}_{k=1,\dots,n}$ and $\{f^k(M^2)\}_{k=1,\dots,n}$ are ε-separated if $dist(f^j(M^1), f^j(M^2) \ge \varepsilon$ for some $j = 1,\dots,n$). Define

$$H_\varepsilon = \limsup_{n\to+\infty} \frac{\ln K(\varepsilon, n)}{n}.$$

Since $K(\varepsilon_1, n) \ge K(\varepsilon_2, n)$ for any $\varepsilon_2 < \varepsilon_1$, it follows that H_ε is non-increasing function of ε. Thus, we may define

$$h = \lim_{\varepsilon\to+0} H_\varepsilon.$$

The topological entropy h may be finite or infinite, however it is finite in the most natural case.

Namely, if Λ is a compact subset of a d-dimensional ball B_d, and f is smooth in B_d, then h is finite. Indeed, let $L = 1 + \sup_{B_d} \|f'\|$. If two orbits $\{f^k(M^1)\}_{k=1,\dots,n}$ and $\{f^k(M^2)\}_{k=1,\dots,n}$ are ε-separated for some small ε, it follows that $dist(f^j(M^1), f^j(M^2)) \ge \varepsilon$, hence, since

$$dist(f^j(M^1), f^j(M^2)) \le \|f'\|^j\ dist(M^1, M^2),$$

we have that

$$dist(M^1, M^2) \geq \varepsilon L^{-n}.$$

Thus, if there is $K(\varepsilon, n)$ of ε-separated orbits in Λ, then there is $K(\varepsilon, n)$ of εL^{-n}-separated points in B_d. It follows that if we surround each of this points by a ball of radius $\frac{1}{2}\varepsilon L^{-n}$, these balls will not intersect. Thus, their summary volume cannot exceed, say, twice the volume of B_d, i.e, it is bounded by a constant C, which does not grow as $\varepsilon \to +0$ or $n \to +\infty$. This gives us

$$K(\varepsilon, n) \leq \frac{C}{(\varepsilon L^{-n})^d},$$

hence

$$\frac{\ln K(\varepsilon, n)}{n} \leq \frac{1}{n} \ln \frac{C}{\varepsilon^d} + d \ln L \to d \ln L,$$

hence

$$H_\varepsilon \leq d \ln L$$

for all ε, hence the topological entropy h satisfies the same inequality and is finite.

Note also that the topological entropy is an invariant of topological conjugacy. Namely, if Λ is a compact invariant set of a map f and Λ' is a compact invariant set of a map g, and if $f|_\Lambda$ is topologically conjugate to $g|_{\Lambda'}$, then

$$h(f|_\Lambda) = h(g|_{\Lambda'}).$$

Indeed, let ψ be the conjugacy, i.e. $\psi \circ f = g \circ \psi$. Since ψ is continuous and Λ is compact, given any ε there exists δ such that if

$$dist(M^1, M^2) \geq \varepsilon$$

for any $M^1 \in \Lambda$, $M^2 \in \Lambda$, then

$$dist(\psi(M^1), \psi(M^2)) \geq \delta.$$

It follows that if two orbits in Λ are ε-separated, then their images by ψ in Λ' are δ-separated. Thus,

$$K(\varepsilon, n; f|_\Lambda) \leq K(\delta, n; g|_{\Lambda'})$$

for every n. This implies

$$H_\varepsilon(f|_\Lambda) \leq H_\delta(g|_{\Lambda'}) \leq h(g|_{\Lambda'}).$$

By taking the limit $\varepsilon \to +0$, we find

$$h(f|_\Lambda) \leq h(g|_{\Lambda'}).$$

Similarly, since ψ^{-1} is uniformly continuous on Λ', we find

$$h(g|_{\Lambda'}) \leq h(f|_\Lambda),$$

which proves the claim.

It is easy to check that for the shift map S on the set of all infinite paths in a graph G formula (18) indeed gives the topological entropy as it is defined for the general case. Thus, Theorem 1 implies

Theorem 3. *For a zero-dimensional compact hyperbolic set Λ of a map f with a Markov partition which corresponds to a graph G, the topological entropy of $f|_\Lambda$ is given by*

$$h = \ln \bar{\lambda}$$

where $\bar{\lambda}$ is the maximum of the absolute value of the eigenvalues of the transition matrix $\mathcal{N}$ of the graph G.

In general, the positivity of the topological entropy is considered to be an equivalent of the "chaoticity" of the dynamics. Theorem 3 shows, in particular, that for hyperbolic sets the topological entropy coincides with the rate at which the number of periodic points grows with period (see (16)). Thus, in this case, chaotic behaviour may be identified with the exponential growth of the number of periodic points. Recall also that all periodic points in the hyperbolic set are unstable. This follows e.g. from the existence of heteroclinics for any pair of periodic orbits in Λ: given any periodic orbit in Λ we have a heteroclinic which leaves an arbitrarily small neighbourhood of the periodic orbits and goes away (tends to another periodic orbit). By the same argument, all the periodic orbits in Λ are unstable for the inverse map as well. Thus, all these periodic orbits are saddle.