CHAPTER III

Chaos and Its Measurement

The theme of the chapter is complicated dynamics or chaos of maps on the line. The first section presents a theorem of Sharkovskii; it proves for certain n and k, that the existence of a periodic orbit of period n forces other orbits of period k. This theorem is not exactly about complicated dynamics, but it does show that if a map f on the line has a period which is not equal to a power of 2, then f has infinitely many different periods. The existence of infinitely many different periods is an indication of the complexity of such a map. This theorem is not used later in the book but is of interest in itself, especially since it is proved using mainly the Intermediate Value Theorem and combinatorial bookkeeping. Sharkovskii's Theorem motivates the treatment of subshifts of finite type which is given in the next section. A subshift of finite type is determined by specifying which transitions are allowed between a finite set of states. Given such a system, it is easy to determine the periods which occur and other aspects of the dynamics. These systems are generalizations of the symbolic dynamics which we introduced for the quadratic map in Chapter II. In Chapter VII and IX, we give further examples of nonlinear dynamical systems which are conjugate to subshifts of finite type. Because we can analyze the subshift of finite type, we determine the complexity of the dynamics of the nonlinear map.

The last few sections of this chapter deal with topics related more directly to chaos. We give a couple of alternative definitions of chaos and introduce various properties which chaotic systems tend to possess. We use this context to introduce the concept of a Liapunov exponent for an orbit. This concept generalizes that of the eigenvalue for a periodic orbit, and associates to an orbit a growth rate of the infinitesimal separation of nearby points. This quantity can be defined even when the map does not have a "hyperbolic structure" like the Cantor set for the quadratic map. This quantity is often used to measure chaos in systems which can be simulated on a computer. In Chapter VIII, we return to define Liapunov exponents in higher dimensions.

3.1 Sharkovskii's Theorem

In Chapter II we showed that the quadratic map $F_\mu(x) = \mu x(1-x)$ has an invariant Cantor set Λ_μ with points of all periods. In this section we study the following question: if a continuous function $f : \mathbf{R} \to \mathbf{R}$ has a point of period n, does it follow that f must have a point of period k? Stated differently, which periods k are forced by which other periods n? The following theorem is very simple to state and has a relatively simple proof.

Theorem 1.1, Li and Yorke (1975). *Assume $f : \mathbf{R} \to \mathbf{R}$ is continuous, and there is a point a such that either (i) $f^3(a) \le a < f(a) < f^2(a)$ or (ii) $f^3(a) \ge a > f(a) > f^2(a)$. Then f has points of all periods.*

REMARK 1.1. Note that it follows if f has a point of period three then it has points of all other periods, hence the title of the Li and Yorke paper, "Period three implies chaos".

REMARK 1.2. There is a more general result of Sharkovskii (1964), which also was proved earlier than the result of Li and Yorke. We prove the simpler result first because the proof is simpler and the lemmas used in the proof of the result of Li and Yorke are needed for the more general result. The treatment of this whole section follows the paper Block, Guckenheimer, Misiurewicz, and Young (1980). Two good general references for these results and other related one dimensional results are Alsedà, Llibre, and Misiurewicz (1993) and Block and Coppel (1992).

We assume the first ordering in the theorem with $f^3(a) \leq a < f(a) < f^2(a)$. (The proof for the other ordering is merely obtained by a reflection in the line.) Let $I_1 = [a, f(a)]$ and $I_2 = [f(a), f^2(a)]$. Then $f(I_1) \supset I_2$ and $f(I_2) \supset I_1 \cup I_2$ as can be seen from the image of the endpoints of the intervals.

Lemma 1.2. *If I and J are closed intervals and $f(I) \supset J$ then there exists a subinterval $K \subset I$ such that $f(K) = J$, $f(\text{int}(K)) = \text{int}(J)$, and $f(\partial K) = \partial J$.*

PROOF. Let $J = [b_1, b_2]$. There exist $a_1, a_2 \in I$ such that $f(a_j) = b_j$. Assume $a_1 < a_2$. (The other case is similar with suprema and infima interchanged.) Let

$$x_1 = \sup\{x : a_1 \leq x \leq a_2 \text{ such that } f(x) = b_1\}.$$

By continuity $f(x_1) = b_1$. Note that $x_1 < a_2$. Next let

$$x_2 = \inf\{x : x_1 \leq x \leq a_2 \text{ such that } f(x) = b_2\}.$$

Then $f(x_2) = b_2$. Thus $f(\{x_1, x_2\}) = \{b_1, b_2\}$. By the definitions of x_1 and x_2, $f((x_1, x_2)) \cap \partial J = \emptyset$. Thus $f(\text{int}([x_1, x_2])) = \text{int}(J) = (b_1, b_2)$. This proves the lemma. □

Definition. An interval I *f-covers an interval* J provided $f(I) \supset J$. We write $I \to J$.

Lemma 1.3. *(a) Assume that there are two points $a \neq b$ with $f(a) > a$ and $f(b) < b$ and $[a, b]$ is contained in the domain of f. Then there is a fixed point between a and b.*

(b) If a closed interval I f-covers itself then f has a fixed point in I.

PROOF. (a) Let $g(x) = f(x) - x$. Then $g(a) > 0$ and $g(b) < 0$. By the Intermediate Value Theorem there is a point c between a and b where $g(c) = 0$ so $f(c) = c$. This result can also be seen graphically by considering the two cases where (i) $a < b$ and (ii) $a > b$. See Figure 1.1.

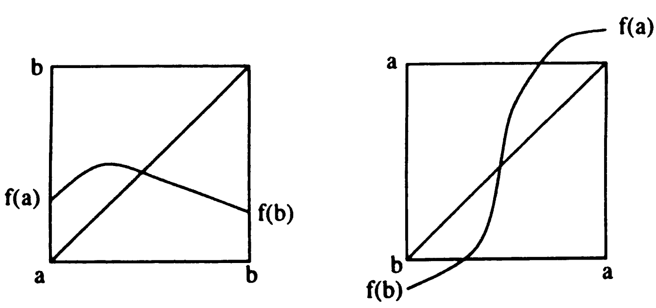

FIGURE 1.1. Fixed Point for Lemma 1.3(a)

(b) By Lemma 1.2, there is an interval $K = [x_1, x_2] \subset I$ with $f(K) = I = [a, b]$. Then either (i) $f(x_1) = a \le x_1$ and $f(x_2) = b \ge x_2$, or (ii) $f(x_1) = b > x_1$ and $f(x_2) = a < x_2$. If the equality holds we are done. Otherwise part (a) applies to prove there is a fixed point. □

Lemma 1.4. *Assume $J_0 \to J_1 \to \cdots \to J_n = J_0$ is a loop with $f(J_k) \supset J_{k+1}$ for $k = 0, \ldots, n-1$.*

(a) Then there exists a fixed point x_0 of f^n with $f^k(x_0) \in J_k$ for $k = 0, \ldots, n$.

(b) Further assume that (i) this loop is not a product loop formed by going p times around a shorter loop of length m where $mp = n$, and (ii) $\text{int}(J_i) \cap \text{int}(J_k) = \emptyset$ unless $J_k = J_i$. If the periodic point x_0 of part (a) is in the interior of J_0 then it has least period n.

REMARK 1.3. Note that the loops that we allow can repeat some intervals as for example $J_0 \to J_1 \to \cdots \to J_{n-2} \to J_0 \to J_0$, or $J_0 \to J_1 \to J_2 \to J_1 \to J_2 \to J_0$. However we do not allow a loop such as $J_0 \to J_1 \to J_0 \to J_1$.

PROOF. (a) We give a proof by induction on j. The induction statement is as follows.

(S_j) There exists a subinterval $K_j \subset J_0$ such that for $i = 1, \ldots, j$, $f^i(K_j) \subset J_i$, $f^i(\text{int}(K_j)) \subset \text{int}(J_i)$, and $f^j(K_j) = J_j$.

By Lemma 1.2 the induction hypothesis is true for $j = 1$.

Assume (S_{k-1}) is true. Thus there exists a K_{k-1}. Then

$$f^k(K_{k-1}) = f(f^{k-1}(K_{k-1})) = f(J_{k-1}) \supset J_k$$

By Lemma 1.2, there exists a subinterval $K_k \subset K_{k-1}$ such that $f^k(K_k) = J_k$ with $f^k(\text{int}(K_k)) = \text{int}(J_k)$. By the induction assumption (S_{k-1}) the other statements of (S_k) are true.

Now using the statement (S_n), we have $f^n(K_n) = J_0$. By Lemma 1.3 f^n has a fixed point x_0 in $K_n \subset J_0$. Because $x_0 \in K_n$, $f^i(x_0) \in J_i$ for $i = 0, \ldots, n$. This proves part (a).

For part (b), since $f^n(\text{int}(K_n)) = \text{int}(J_0)$, if $x_0 \in \text{int}(J_0)$ then $x_0 \in \text{int}(K_n)$ and $f^i(x_0) \in \text{int}(J_i)$ for $i = 1, \ldots, n$. Because the loop is not a product, x_0 must have period n. □

PROOF OF THEOREM 1.1.

We assume the first case where $f(a) = b > a$, $f^2(a) = f(b) = c > f(a) = b$, and $f^3(a) = f(c) \le a$. Let $I_1 = [a, b]$ and $I_2 = [b, c]$. Then I_1 f-covers I_2 and I_2 f-covers both I_1 and I_2.

First $F(I_2) \supset I_2$ so there is a fixed point by Lemma 1.3.

Next we show that f has a point of period n for any $n \ge 2$. Take the loop of length n with one interval being I_1 and $n-1$ intervals being repeated copies of I_2: $I_1 \to I_2 \to I_2 \to \cdots \to I_2 \to I_1$. By Lemma 1.4, there exists an $x_0 \in I_1$ such that $f^n(x_0) = x_0$ and $f^j(x_0) \in I_2$ for $j = 1, \ldots, n-1$. If there were a k with $1 \le k < n$ such that $f^k(x_0) = x_0$, then we would have $x_0 = f^k(x_0) \in I_2$. Thus we would have $x_0 \in I_1 \cap I_2 = \{b\}$. We now show that $x_0 = b$ is impossible. The argument is slightly different for $n = 2$ and $n \ge 3$. In the case when $n = 2$, $f^2(b) = f^2(x_0) = x_0 = b$, contradicting $f^2(b) = f^3(a) \le a$. In the case when $n \ge 3$, we must have $f^2(b) = f^2(x_0) \in I_2$ contradicting $f^2(b) = f^3(a) \le a$. This contradiction shows that $f^j(x_0) \ne x_0$ for $1 \le j < n$, and x_0 has period n. □

Definition. In order to state the result of Sharkovskii we need to introduce a new ordering on the positive integers using the symbol $\triangleright$ called the *Sharkovskii ordering*. First the odd integers greater than one are put in the backward order:

$$3 \triangleright 5 \triangleright 7 \triangleright 9 \triangleright 11 \triangleright \cdots.$$

Next, all the integers which are two times an odd integer are added to the ordering, and then the odd integers times increasing powers of two:

$$3 \triangleright 5 \triangleright 7 \triangleright \cdots \triangleright 2 \cdot 3 \triangleright 2 \cdot 5 \triangleright \cdots \triangleright 2^2 \cdot 3 \triangleright 2^2 \cdot 5 \triangleright \cdots$$
$$\triangleright 2^n \cdot 3 \triangleright 2^n \cdot 5 \triangleright \cdots \triangleright 2^{n+1} \cdot 3 \triangleright 2^{n+1} \cdot 5 \triangleright \cdots .$$

Finally, all the powers of two are added to the ordering in decreasing powers:

$$3 \triangleright 5 \triangleright \cdots \triangleright 2^n \cdot 3 \triangleright 2^n \cdot 5 \triangleright \cdots \triangleright \cdots \triangleright 2^{n+1} \triangleright 2^n \triangleright \cdots \triangleright 2^2 \triangleright 2 \triangleright 1.$$

We have now given an ordering between all positive integers. This ordering seems strange but it turns out to the be ordering which expresses which periods imply which other periods as given in the Theorem of Sharkovskii (Sharkovskii, 1964).

Theorem 1.5 (Sharkovskii). *Let $f : I \subset \mathbb{R} \to \mathbb{R}$ be a continuous function from an interval I into the real line. Assume f has a point of period n and $n \triangleright k$. Then f has a point of period k. (By period we mean least period.)*

Until the proof of the theorem is complete, f is assumed to be a continuous function from I to $\mathbb{R}$ as given in the statement. The proof of the theorem involves finding intervals which f-cover each other in certain ways. In order to express these ideas we introduce the following definition of a type of graph.

Definition. Let $\mathcal{A} = \{I_1, \ldots, I_s\}$ be a partition of I into closed intervals with disjoint nonempty interiors. An *$\mathcal{A}$-graph of f* is a directed graph with vertices given by the I_j and a directed edge from I_j to I_k if I_j f-covers I_k. It is also called the *graph for the partition.* See Figures 1.3 and 1.4 for examples.

Example 1.1. Let f have a graph as indicated in Figure 1.2 with three intervals, I_1, I_2, I_3. Then I_1 f-covers I_2, I_2 f-covers I_1 and I_2, and I_3 f-covers I_1, I_2, and I_3. Thus the graph for the partition is as given in Figure 1.3.

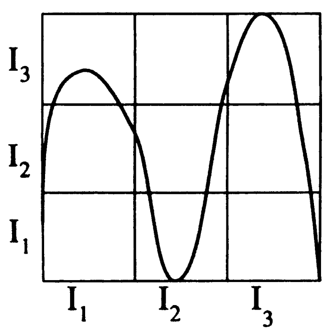

FIGURE 1.2. Graph of the Function in Example 1.1

REMARK 1.4. We first consider the case where n is an odd integer for which (i) $n > 1$ and (ii) f has a point x of period n and f has no points of odd period k with $1 < k < n$ (i.e., $k \triangleright n$.) To prove Sharkovskii's Theorem in this case, Peter Stefan had the idea to

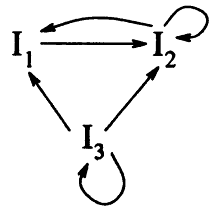

FIGURE 1.3. Graph for the Partition in Example 1.1

prove the existence of an orbit with a special pattern on the line: let x_1 be an n-periodic point such that

$$x_n < x_{n-2} < \cdots < x_3 < x_1 < x_2 < x_4 < \cdots < x_{n-1}$$

where $x_j = f^{j-1}(x_1)$. (The reflection of this ordering is just as good.) A periodic point with such an ordering of its orbit on the line is called a *Stefan cycle*. Lemma 1.6 proves that indeed such an orbit does exist. Given such an orbit, let $I_1 = [x_1, x_2]$, $I_2 = [x_3, x_1]$, $I_3 = [x_2, x_4]$, $I_{2j} = [x_{2j+1}, x_{2j-1}]$, and $I_{2j-1} = [x_{2j-2}, x_{2j}]$ for $j = 2, \ldots, (n-1)/2$. Because of the nature of the orbit, (i) I_1 f-covers I_1 and I_2, (ii) I_j f-covers I_{j+1} for $2 \leq j \leq n-2$, and (iii) I_{n-1} f-covers all the I_j for j odd. Thus the existence of such a special type of orbit proves that the $\mathcal{A}$-graph of f contains a subgraph of the form given in Figure 1.4. This subgraph is called a *Stefan graph*. Applying the lemmas above to this Stefan graph can prove the existence of all the periodic implied by n in the Sharkovskii ordering.

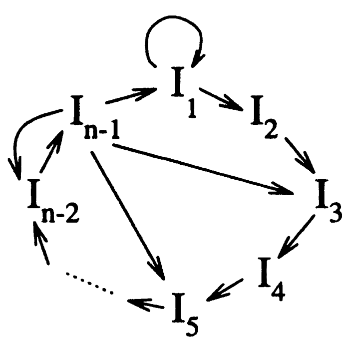

FIGURE 1.4. Subgraph for the Partition in Lemma 1.6

We now turn to the lemma and its proof.

Lemma 1.6. *Assume n is an odd integer with $n > 1$. Assume that f has a point x of period n and f has no points of odd period k with $1 < k < n$ (i.e., $k \triangleright n$.) Let $J = [\min \mathcal{O}(x), \max \mathcal{O}(x)]$. Let $\mathcal{A}$ be the partition of J by the elements of $\mathcal{O}(x)$. Then the $\mathcal{A}$-graph of f contains a subgraph of the following form: The $I_1, \ldots, I_{n-1}$ can be numbered with all the intervals having disjoint interiors such that (i) I_1 f-covers I_1 and I_2, (ii) I_j f-covers I_{j+1} for $2 \leq j \leq n-2$, and (iii) I_{n-1} f-covers all the I_j for j odd. See Figure 1.4.*

PROOF. Let $\mathcal{O}(x) = \{z_1, z_2, \ldots, z_n\}$ where the z_j are ordered as on the line, $z_1 < z_2 < \cdots < z_n\}$. Then, $f(z_n) < z_n$ because $f(z_n)$ is one of the other z_j. Similarly, $f(z_1) > z_1$.

Let $a = \max\{y \in \mathcal{O}(x) : f(y) > y\}$. Then, $a \neq z_n$. Let b be the next larger than a among $\mathcal{O}(x)$ in terms of the ordering of the real line. Let $I_1 = [a, b] \in \mathcal{A}$. We show that this I_1 can be used in the statement of the lemma.

There is a sequence of small steps which we state as claims. First we need to show that I_1 covers itself (Claim 1), and eventually covers all of J (Claim 2). Claim 4 shows that there is a shortest loop with distinct intervals $I_1 \to I_2 \to \cdots \to I_{n-1} \to I_1$. Claims 5 and 6 complete showing that these intervals are situated on the line and behave as claimed.

Claim 1. *The image of I_1 covers itself, $f(I_1) \supset I_1$.*

PROOF. We know that $f(a) > a$ so $f(a) \geq b$. Also since $b > a$, $f(b) < b$ so $f(b) \leq a$. Therefore $f(I_1) \supset I_1$ as claimed. □

Claim 2. *The $(n-2)$ image of I_1 covers the whole interval J, $f^{n-2}(I_1) \supset J$.*

PROOF. Since $f(I_1) \supset I_1$, $f^{k+1}(I_1) \supset f^k(I_1)$, so the iterates are nested. The number of points in $\mathcal{O}(x) \setminus \{a, b\}$ is $n - 2$, so $z_n \in f^k(I_1)$ for some $0 \leq k \leq n - 2$. By the nested property, $z_n \in f^{n-2}(I_1)$. Similarly $z_1 \in f^{n-2}(I_1)$. Since I_1 is connected, $f^{n-2}(I_1) \supset [z_1, z_n] = J$. □

Claim 3. *There exists a $K_0 \in \mathcal{A}$ with $K_0 \neq I_1$ such that $f(K_0) \supset I_1$.*

PROOF. This proof uses the fact that n is odd, so there are more elements of $\mathcal{O}(x)$ on one side of $\mathrm{int}(I_1)$ than the other. Call $\mathcal{P}$ the elements of $\mathcal{O}(x)$ on the side of $\mathrm{int}(I_1)$ with more elements. There is some $y_1, y_2 \in \mathcal{P}$ with $f(y_1) \in \mathcal{P}$ and $f(y_2) \in \mathcal{O}(x) \setminus \mathcal{P}$. Take adjacent points y_1 and y_2 with iterates as above. Let K_0 be the interval from y_1 to y_2. Then $f(K_0) \supset I_1$ and $K_0 \neq I_1$ as claimed. □

Claim 4. *There is a loop $I_1 \to I_2 \to \cdots \to I_k \to I_1$ with $I_2 \neq I_1$. The shortest such loop with $k \geq 2$ has $k = n - 1$.*

PROOF. Let K_0 be as in Claim 3, so $f(K_0) \supset I_1$. By Claim 2, $f^{n-2}(I_1) \supset K_0$. There are only $n - 1$ distinct intervals in $\mathcal{A}$ so there exists such a loop with $2 \leq k \leq n - 1$.

Now assume the smallest k that works satisfies $2 \leq k < n - 1$ and we get a contradiction. Since this is the shortest loop, none of the intervals can be repeated or it could be shortened. Either k or $k + 1$ is odd. Let $m = k$ or $k + 1$ be this odd integer, so $1 < m < n$. Use the loop with m intervals given by $I_1 \to I_2 \to \cdots \to I_k \to I_1$ or $I_1 \to I_2 \to \cdots \to I_k \to I_1 \to I_1$ depending on whether $m = k$ or $m = k + 1$. By Lemma 1.4(a) there is a point z with $f^m(z) = z$. The point z can not be on the boundary of the interval because these points have period n which is greater than m. Thus z has least period m by Lemma 1.4(b). Since m is odd this contradicts the assumption on n in the Lemma. This contradiction proves that $k = n - 1$. □

For the rest of the proof we fix $I_1, I_2, \ldots, I_{n-1}$ as in Claim 4.

Claim 5. *(a) If $f(I_j) \supset I_1$ then $j = 1$ or $n - 1$.*
(b) For $j > i + 1$ there is no directed edge from I_i to I_j in the graph.
(c) The interval I_1 f-covers only I_1 and I_2.

PROOF. Part (a) follows from Claim 4. Parts (b) and (c) follow because the loop is the shortest possible. □

Claim 6. *Either (i) the ordering (in terms of the real line) of the intervals I_j in the loop of Claim 4 is $I_{n-1} \leq I_{n-3} \leq \cdots \leq I_2 \leq I_1 \leq I_3 \leq \cdots \leq I_{n-2}$ and the order of the* *orbit is $f^{n-1}(a) < f^{n-3}(a) < \cdots < f^2(a) < a < f(a) < f^3(a) < \cdots < f^{n-2}(a)$ or (ii) both of these orderings are exactly reversed.*

PROOF. Let $I_1 = [a, b]$. The interval I_1 f-covers only I_1 and I_2 so they must be next to each other. Assume that $I_2 \leq I_1$. (The other possibility gives the reverse order mentioned in the claim.) Then it must be that $f(a) = b$ and $f(b)$ is the left endpoint of I_2.

Next, $f(\partial I_2) = \partial I_3$. Since one of these endpoints is $f(a) = b$ which is above $\text{int}(I_1)$ both endpoints of I_3 must be above $\text{int}(I_1)$. Also because of Claim 5a (I_2 does not f-cover I_1) and 5b (I_2 does not f-cover I_j for $j > 3$) I_3 must be adjacent to I_1.

Continue the argument by induction. For $k < n - 1$, since I_k does not f-cover I_1 and I_k does not f-cover I_j for $j > k + 1$, I_{k+1} must be adjacent to I_{k-1}. This covers all the intervals in the claim.

Note that we have also shown the ordering on the orbit as stated in the claim. □

Claim 7. *The interval I_{n-1} f-covers all the I_j for j odd.*

PROOF. Note that $I_{n-1} = [f^{n-1}(a), f^{n-3}(a)]$. Then $f(f^{n-1}(a)) = f^n(a) = a$. Also $f^{n-3}(a) \in I_{n-3}$ so $f(f^{n-3}(a)) = f^{n-2}(a) \in I_{n-2}$ is the far right endpoint of J (the largest element in the orbit $\mathcal{O}(x)$). Thus $f(I_{n-1}) \supset [a, f^{n-2}(a)] = I_1 \cup I_3 \cup \cdots \cup I_{n-2}$. We have proved the claim. □

All the claims together prove Lemma 1.6. □

Proposition 1.7. *Theorem 1.5 is true if n is odd and maximal in the ordering for which the theorem is true.*

PROOF. Take k with $n \rhd k$. There are two cases: (a) k is even and $k < n$ and (b) $k > n$ with k either even or odd.

Case a. *The integer k is even and $k < n$.*

PROOF. Consider the loop of length k given by $I_{n-1} \to I_{n-k} \to I_{n-k+1} \to \cdots \to I_{n-1}$. By Lemma 1.4(a) there is a $x_0 \in I_{n-1}$ with $f^k(x_0) = x_0$. The point x_0 can not be an endpoint because the endpoints have period n. Therefore x_0 has period k. □

Case b. *The integer $k > n$ with k either even or odd.*

PROOF. Consider the loop of length k given by $I_1 \to I_2 \to \cdots \to I_{n-1} \to I_1 \to I_1 \to \cdots \to I_1$. Again by Lemma 1.4(a) there is a $x_0 \in I_1$ with $f^k(x_0) = x_0$. If $x_0 \in \partial I_1$ then x_0 has period n. Thus n divides k, so $k \geq 2n \geq n + 3$. Also since $f^n(x_0) \in I_1$ the iterate $f^{n+1}(x_0) \notin I_1$ which contradicts the conclusion of Lemma 1.4(a). Therefore $x_0 \notin \partial I_1$, and by Lemma 1.4(b), x_0 has period k. This completes the proof of Case b and Proposition 1.7. □

The first step in proving the result for other values of n proves the existence of a point of period two whenever there is a point of even period.

Lemma 1.8. *If f has a point of even period then it has a point of period two.*

PROOF. Let n be the smallest integer greater than one in the usual ordering of the integers (not the Sharkovskii ordering) such that f has a point of period n. If n is odd then we are done by Proposition 1.7. Therefore we can assume that n is even. Let a, $I_1 = [a, b]$, and $J = [\min \mathcal{O}(a), \max \mathcal{O}(a)] = [A, B]$ be as before. In the proof of Lemma 1.6 we only used the fact that n is odd to show that there exists a $K_0 \in \mathcal{A}$ with $K_0 \neq I_1$ and $f(K_0) \supset I_1$.

First assume there is such a K_0. There is a minimal cycle as in Claim 4 with $2 \leq k \leq n-1$. As before, I_k covers all the I_j on the other side. Thus $I_{n-1} \to I_{n-2} \to I_{n-1}$ is a cycle of length two, and there is a point of period two.

Next assume there is no $K_0 \in \mathcal{A}$ with $K_0 \neq I_1$ and $f(K_0) \supset I_1$. It follows that (i) all the points $x_j \in \mathcal{O}(a)$ with $x_j \leq a$ have $f(x_j) \geq b$ and (ii) all the points $x_j \in \mathcal{O}(a)$ with $x_j \geq b$ have $f(x_j) \leq b$. Since some points in $\mathcal{O}(a)$ are mapped to b and B, both $b, B \in f([A,a])$ and so $f([A,a]) \supset [b,B]$. Similarly, $f([b,B]) \subset [A,a]$. Then $[A,a] \to [b,B] \to [A,a]$ is a cycle of length two. The intervals are disjoint so there must be a point of period two. □

The proof of Sharkovskii's Theorem now splits into the following cases.

Case 1: n is odd and maximal in the Sharkovskii ordering and $n \rhd k$.
Case 2: $n = 2^m$ and $n \rhd k$.
Case 3: $n = 2^m p$ with $p > 1$ odd, $m \geq 1$, n is maximal in the Sharkovskii ordering, and $n \rhd k$.

Case 1 is proved above in Proposition 1.7. We split Case 2 up into subcases and prove it next.

Case 2: $n = 2^m$ and $n \rhd k$ so $k = 2^s$ with $0 \leq s < m$.
Case 2a: $s = 0$, i.e., f has a fixed point.
Case 2b: $s = 1$.
Case 2c: $s > 1$.

PROOF OF CASE 2a. We can define a and b as before with $f(a) \geq b$ and $f(b) \leq a$. Therefore $f([a,b]) \supset [a,b]$ and f has a fixed point. □

Case 2b follows from Lemma 1.8.

PROOF OF CASE 2c. Let $g = f^{k/2} = f^{2^{s-1}}$. The map g has a point of period 2^{m-s+1} with $m - s + 1 \geq 2$. Lemma 1.8 proves that g has a point x_0 of period 2. So $x_0 = g^2(x_0) = f^k(x_0)$ and $x_0 \neq g(x_0) = f^{k/2}(x_0)$. Thus the period of x_0 for f is 2^t for some $t \leq s$. If $t < s$ then x_0 is fixed by g which is impossible. Therefore $t = s$ and x_0 is a point of period $2^s = k$. □

We also split Case 3 up into subcases.

Case 3: $n = 2^m p$ with $p > 1$ odd, $m \geq 1$, n is maximal in the Sharkovskii ordering for f, and $n \rhd k$.
Case 3a: $k = 2^s q$ with $s \geq m+1$ and $1 \leq q$ and q odd.
Case 3b: $k = 2^s$ with $s \leq m$.
Case 3c: $k = 2^m q$ with q odd and $q > p$.

We leave the proof of these cases to the exercises. See Exercises 3.2 - 3.4. This completes the proof of Theorem 1.5 (Sharkovskii's Theorem). □

3.1.1 Examples for Sharkovskii's Theorem

There are examples of maps with exactly the orbits implied by the Sharkovskii ordering. First consider the case where the maximal period in the ordering is odd.

Example 1.2. Let $n > 3$ be odd. (If $n = 3$ there are points of all periods and there is nothing to prove.) Let x_1 be a point which is a Stefan cycle for n. Make the graph be piecewise linear connecting the adjacent points $(x_j, f(x_j))$ on the graph by straight line segments. Let $I_1, \ldots, I_{n-1}$ be the intervals as in the proof. See Figure 1.5.

We claim that such a map does not have a point of odd period k with $1 < k < n$. Assume that x is a periodic point with period different than n. If any iterate of x hits one of the endpoints of an I_j then either x has period n or is not periodic and so this can not happen. Thus $f^j(x) \in \text{int}(I_{i(j)})$ for each j. Because the $\mathcal{A}$-graph is exactly the

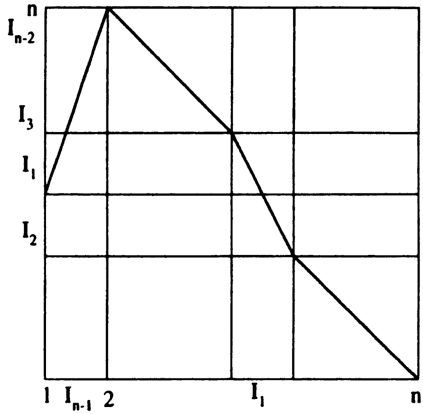

FIGURE 1.5. Example 1.2

subgraph proved to exist by Lemma 1.6, the length of cycles in the graph are exactly those k which are implied by n in the Sharkovskii ordering. Also the graph over I_1 has slope -2. There is a fixed point in I_1 and all other points must leave I_1 and enter I_2. Thus all orbits passing through I_1 are either fixed points or have periods at least $n-1$. Other orbits have to have the same period as the period of the cycle of intervals (since the orbit must pass through the interiors). Thus the possible periods of periodic points are exactly those implied by n in the Sharkovskii ordering. In particular there are no points of odd period with $1 < k < n$.

Definition. To get other examples with certain periods we introduce the doubling operator. Let $I = [0,1]$. Assume $f : I \to I$ is a continuous map. We denote the periods of the orbits of f by $\mathcal{P}(f)$. Now we define the *double of* f, $\mathcal{D}(f) = g$, by

$$g(x) = \begin{cases} \frac{2}{3} + \frac{1}{3}f(3x) & \text{for } 0 \le x \le \frac{1}{3} \\ [2 + f(1)](\frac{2}{3} - x) & \text{for } \frac{1}{3} \le x \le \frac{2}{3} \\ x - \frac{2}{3} & \text{for } \frac{2}{3} \le x \le 1. \end{cases}$$

See Figure 1.6. It is easily checked that g is continuous.

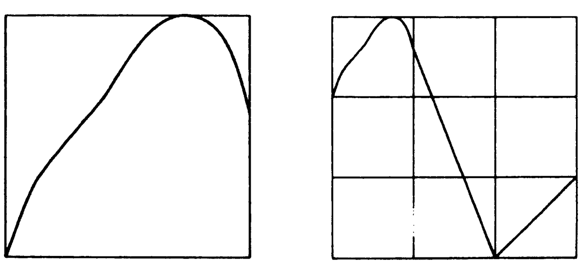

FIGURE 1.6. The original map f is in the left figure and the double g is in the right figure

The next proposition relates the periods of f with the periods of the double of f; this result justifies the use of the name "double" for this construction.

Proposition 1.9. *The set of all periods of g, $\mathcal{P}(g)$, are related to the set of all periods of f, $\mathcal{P}(f)$, by $\mathcal{P}(g) = 2\mathcal{P}(f) \cup \{1\}$. Moreover g has exactly one repelling fixed point, and for each n g has the same number of orbits of period $2n$ as f has of period n and their stability is the same.*

PROOF. Let $I_1 = [0, 1/3]$, $I_2 = (1/3, 2/3)$, and $I_3 = [2/3, 1]$. Because $g(I_2) \supset I_2$, g has a fixed point x_1 in I_2. Because the absolute value of the slope of g in I_2 is at least 2, there is exactly one fixed point in I_2 and it is repelling. Also any point in I_2 other than the fixed point has an orbit which leaves I_2. Because $[g(I_1) \cup g(I_3)] \cap I_2 = \emptyset$, none of the points in $I_2 \setminus \{x_1\}$ can be periodic. If $x \in I_1$ then $g^2(x) = g(2/3 + f(3x)/3) = f(3x)/3 \in I_1$. Thus for x to be periodic, its period must be even, $2k$. But by induction, $g^{2k}(x) = f^k(3x)/3$ for $k \geq 1$. Thus $g^{2k}(x) = x$ if and only if $f^k(3x) = 3x$. We have shown that these periods of g are exactly twice the periods of f. Moreover, since $g'(t) = 1$ on I_3, for a point x of period $2k$ for g, $(g^{2k})'(x) = (f^k)'(3x)$ so the two orbits have the same stability type. The periodic points of g in I_3 are the same because they are on the orbits described above. This proves the proposition. □

Example 1.3. Let $f(x) \equiv 1/3$ for $x \in [0, 1]$. The only periodic point of f is a fixed point. Let $f_1 = \mathcal{D}(f)$. By the above proposition the periods of f_1, $\mathcal{P}(f_1)$, are $\{1, 2\}$. Also, f_1 has one repelling fixed point and one attracting orbit of period 2. By induction, if $f_n = \mathcal{D}^n(f)$ then the periods of f_n, $\mathcal{P}(f_n)$, are $\{1, 2, \ldots, 2^n\}$, and f_n has one repelling periodic orbit of period 2^j for $1 \leq j < n$ and one attracting periodic orbit of period 2^n. Finally let $f_\infty(x) = \lim_{n\to\infty} f_n(x)$. We leave to an exercise to prove that f_∞ is continuous and $\mathcal{P}(f_\infty) = \{1, 2, \ldots, 2^n, \ldots\}$, i.e., f_∞ has repelling periodic points of periods 2^n for all n and no other periods. See Exercise 3.8.

We also leave to the exercises the fact that if $n = 2^m p$ for $1 < p$, p odd, and $m \geq 1$, then there is a map f for which $\mathcal{P}(f) = \{k : n \rhd k\}$. See Exercise 3.5.

3.2 Subshifts of Finite Type

In the proof of Sharkovskii's Theorem we considered graphs where intervals f-covered each other forming a graph as given in Figure 2.1.

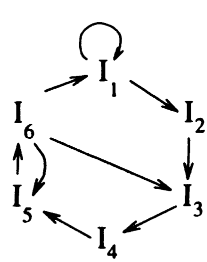

FIGURE 2.1. Graph of Partition in Sharkovskii's Theorem

With this graph, a point in I_1 can go to I_1 or I_2; a point in I_2 can go to I_3; a point in I_3 can go to I_4; a point in I_4 can go to I_5; a point in I_5 can go to I_6; a point in I_6 can go to I_1, I_3, or I_5. Paths in the graph correspond to allowable orbits of points. We can look at only the labeling of the intervals (as we did for the quadratic map $f_\mu(x) = \mu x(1-x)$

for $\mu > 2 + 5^{1/2}$) and consider sequences $\mathbf{s} = s_0 s_1 s_2 \ldots$ where 1 can be followed by 1 or 2; 2 can only be followed by 3; 3 can only be followed by 4; 4 can only be followed by 5; 5 can only be followed by 6; can 6 can be followed by 1, 3, or 5. All other adjoining combinations are not allowed. Thus a sequence like $634561123456\ldots$ is allowed.

Definition. Instead of looking at the graph, we can define a *transition matrix* to be a matrix $A = (a_{ij})$ such that (i) $a_{ij} = 0, 1$ for all i and j, (ii) $\sum_j a_{ij} \geq 1$ for all i, and (iii) $\sum_i a_{ij} \geq 1$ for all j. Given a graph of the type in Sharkovskii's Theorem, we can form a transition matrix A by letting $a_{ij} = 0$ if the transition from i to j is not allowed (there is no arrow in the graph from I_i to I_j) and $a_{ij} = 1$ if the transition from i to j is allowed (there is an arrow in the graph from I_i to I_j). The assumption that $\sum_j a_{ij} \geq 1$ for every i means that it is possible to go to some interval from I_i; the assumption that $\sum_i a_{ij} \geq 1$ for every j means that it is possible to get back to I_j from some interval. In the above graph the transition matrix is

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

Definition. Let Σ_n be the space of all (one-sided) sequences with symbols in the set $\{1, 2, \ldots, n\}$ as defined in Section 2.5, and $\sigma : \Sigma_n \to \Sigma_n$ be the shift map given by $\sigma(\mathbf{s}) = \mathbf{t}$ where $t_k = s_{k+1}$. This space has a metric as defined before.

Given an n by n transition matrix A let

$$\Sigma_A = \{\mathbf{s} \in \Sigma_n : a_{s_k s_{k+1}} = 1 \text{ for } k = 0, 1, 2, \ldots\}.$$

This space Σ_A is made up of the allowable sequences for A. Let $\sigma_A = \sigma|\Sigma_A$. The following proposition shows that σ_A acting on a sequence in σ_A gives another sequence in σ_A. The map $\sigma_A : \Sigma_A \to \Sigma_A$ is called the ***subshift of finite type for the matrix A***. The following proposition also shows that Σ_A is closed.

Proposition 2.1. *(a) The subset Σ_A is closed in Σ_n.*

(b) The map σ_A leaves Σ_A invariant, $\sigma_A(\Sigma_A) = \Sigma_A$.

PROOF. (a) By using cylinder sets it is easily seen that Σ_A is closed.

(b) If $\mathbf{s} \in \Sigma_A$ and $\mathbf{t} = \sigma_A(\mathbf{s})$ then it follows directly that all the transitions in $\mathbf{t}$ are allowed so $\mathbf{t} \in \Sigma_A$. On the other hand, if $\mathbf{t} \in \Sigma_A$ then there is some s_0 such that $a_{s_0 t_0} = 1$ by the standing assumptions on A. Let $s_k = t_{k-1}$ for $k \geq 1$. Then $\mathbf{s} \in \Sigma_A$ and $\sigma_A(\mathbf{s}) = \mathbf{t}$. □

Definition. In general, a subset $S \subset \Sigma_n$ is called a *subshift* provided that it is closed and invariant by the shift map σ. The following example gives a subshift which is not of finite type.

Example 2.1. Let S be the subset of Σ_2 consisting of all strings s such that between any two 2's in the string $\mathbf{s}$ there are an even number of 1's: i.e., if $s_j = 2 = s_k$ with $j < k$ then there are an even number of indices i with $j < i < k$ for which $s_i = 1$. This allows the string $\mathbf{s}$ to start with an odd number of 1's, and $\mathbf{s}$ can have an infinite tail of all 1's or all 2's. A direct check shows that S is closed and invariant under the shift map. Because the number of 1's between two 2's can be an arbitrary even number, S is not a subshift of finite type.

The next thing we want to do is count the number of periodic orbits in Σ_A. A string which has period k for σ_A keeps repeating the first k symbols that appear in its string, e.g. 124212421242... has period 4. Therefore it is helpful to look at finite strings of symbols which are called *words*. Therefore 1242 is a word of length 4. Given a transition matrix A, a word $\mathbf{w} = (w_0, \dots, w_{k-1})$ is called *allowable* provided the transition from w_{j-1} to w_j is allowable for $j = 1, \dots, k$, i.e., $a_{w_{j-1},w_j} = 1$ for $j = 1, \dots, k$. As a first step (the induction step) to determine the number of k-periodic points for σ_A, we prove the following lemma about the number of words of length $k+1$ which start at any symbol i and end at the symbol j.

Lemma 2.2. *Assume that the ij entry of A^k is p, $(A^k)_{ij} = p$. Then there are p allowable words of length $k+1$ starting at i and ending at j, i.e., words of the form $is_1s_2\dots s_{k-1}j$.*

PROOF. We prove the result by induction on k. Let $\operatorname{num}(k,i,j)$ be the number of words of length $k+1$ starting at i and ending at j. This result is certainly true for $k = 1$ where $\operatorname{num}(1,i,j)$ is either zero or one depending on whether there is an allowable transition from i to j or not. Now assume the lemma is true for $k-1$ for all choices of i and j. By matrix multiplication

$$\begin{aligned}
(A^k)_{ij} &= \sum_{s_1,s_2,\dots,s_{k-1}} a_{is_1}a_{s_1s_2}\cdots a_{s_{k-1}j} \\
&= \sum_{s_{k-1}} \Big(\sum_{s_1,s_2,\dots,s_{k-2}} a_{is_1}a_{s_1s_2}\cdots a_{s_{k-2}s_{k-1}} \Big) a_{s_{k-1}j} \\
&= \sum_{s_{k-1}} \operatorname{num}(k-1,i,s_{k-1}) a_{s_{k-1}j} \\
&= \operatorname{num}(k,i,j).
\end{aligned}$$

The last equality is true because if $a_{s_{k-1}j} = 0$ then these words from i to s_{k-1} do not contribute to the count of the words from i to j. On the other hand, if $a_{s_{k-1}j} = 1$ then each of these words contributes one word to the words from i to j. □

Corollary 2.3. *The number of fixed points of σ_A^k is equal to the trace of A^k.*

PROOF. This follows because $\#\operatorname{Fix}(\sigma_A^k|\Sigma_A) = \sum_i \operatorname{num}(k,i,i) = \sum_i (A^k)_{ii} = \operatorname{tr}(A^k)$. □

Definition. An n by n matrix of 0's and 1's is called *reducible* provided that there is a pair i, j with $(A^k)_{ij} = 0$ for all $k \geq 1$. An n by n matrix of 0's and 1's is called *irreducible* provided that for each $1 \leq i, j \leq n$ there exists a $k = k(i,j) > 0$ such that $(A^k)_{ij} > 0$, i.e., there is an allowable sequence from i to j for every pair of i and j. The matrix A is called *positive* provided $A_{ij} > 0$ for all i and j and is called *eventually positive* provided there there exists a k which is independent of i and j such that $(A^k)_{ij} > 0$ for all i and j. Thus, both positive and eventually positive matrices are irreducible.

Example 2.2. (a) The following transition matrix is reducible because it is not possible to get from 3 to 1:

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

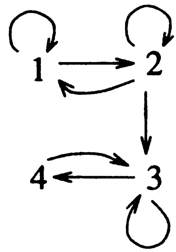

FIGURE 2.2. Graph for Partition in Example 2.2(a)

Its graph is given in Figure 2.2.

(b) The following transition matrix is irreducible:

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

Its graph is given in Figure 2.3.

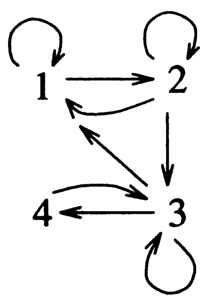

FIGURE 2.3. Graph for Partition in Example 2.2(b)

We often want to exclude the case when A corresponds to a permutation of symbols. A permutation is defined to be a transition matrix where the sum of each row is equal to one and the sum of each column is also equal to one. An example of a permutation is given by

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

Its graph is given in Figure 2.4.

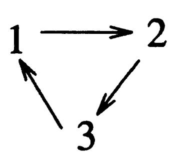

FIGURE 2.4. Graph for Permutation on Three Symbols

The following lemma characterizes permutations in terms of row sums only.

Lemma 2.4. *Let A be a transition matrix. Then A is a permutation matrix if and only if $\sum_j a_{ij} = 1$ for all i.*

PROOF. If every row sum, $\sum_j a_{ij} = 1$ for all i, then $\sum_{i,j} a_{ij} = n$. Since a transition matrix has $\sum_i a_{ij} \geq 1$ for every j, it follows that $\sum_i a_{ij} = 1$ for every j. This shows A is a permutation matrix. The converse is clear. □

In the next proposition, we show that a subshift is irreducible if and only if the shift map has a dense orbit. As a preliminary step, we prove the following lemma about the dense orbit.

Lemma 2.5. *Let A be a transition matrix. Assume that the point $\mathbf{s}^*$ has a dense orbit in Σ_A for the shift map σ_A and $\mathbf{s}^*$ is not a periodic point. (This last assumption means that A is not a permutation matrix.) Then for any $k > 1$, $\sigma_A^k(\mathbf{s}^*)$ has a dense orbit.*

PROOF. It is clear that $\sigma_A^{k+j}(\mathbf{s}^*)$ is dense in $\Sigma_A \setminus \{\sigma_A^i(\mathbf{s}^*) : 0 \leq i < k\}$. Thus we only need to show that this orbit accumulates on $\{\sigma_A^i(\mathbf{s}^*) : 0 \leq i < k\}$. It is clearly sufficient to prove that $\sigma_A^{k+j}(\mathbf{s}^*)$ accumulates on $\mathbf{s}^*$ by taking a higher iterate to get near the other $\sigma_A^i(\mathbf{s}^*)$. Rather than look at $\mathbf{s}^*$, we show that $\mathbf{s}^*$ has a preimage $\mathbf{t}^*$ and that $\sigma_A^{k+j}(\mathbf{s}^*)$ accumulates on $\mathbf{t}^*$.

First, we show $\mathbf{s}^*$ has a preimage. By assumption (iii) for a transition matrix, there is an element t_0 that can make a transition to $s_0^* \equiv t_1$, so there is a $\mathbf{t}^* \in \sum_A$ with $\sigma_A(\mathbf{t}^*) = \mathbf{s}^*$. If $\mathbf{t}^*$ were on the forward orbit of $\mathbf{s}^*$ then $\mathbf{s}^*$ would be periodic which is not allowed. Therefore $\mathbf{t}^*$ is not on the forward orbit of $\mathbf{s}^*$, and so $\sigma_A^j(\mathbf{s}^*) \neq \mathbf{t}^*$ for $0 \leq j \leq k$.

Since $\sigma_A^j(\mathbf{s}^*)$ is dense everywhere in Σ_A, and $\sigma_A^j(\mathbf{s}^*) \neq \mathbf{t}^*$ for $0 \leq j \leq k$, it follows that $\sigma_A^{k+j}(\mathbf{s}^*)$ must come closer to $\mathbf{t}^*$ than the finite set of points $\{\sigma_A^j(\mathbf{s}^*) : 0 \leq j \leq k\}$. Therefore $\sigma_A^j \circ \sigma_A^k(\mathbf{s}^*)$ comes arbitrarily near $\mathbf{t}^*$. This completes the proof that the forward orbit of $\sigma_A^k(\mathbf{s}^*)$ is dense in Σ_A. □

Proposition 2.6. *Let $A = (a_{i,j})$ be a transition matrix. Then the following are equivalent:*

(a) *A is irreducible, and*
(b) *σ_A has a dense forward orbit in Σ_A.*

PROOF. For a finite word $\mathbf{w}$, let $b(\mathbf{w})$ be the first letter of $\mathbf{w}$ (beginning of $\mathbf{w}$), and $e(\mathbf{w})$ be the last letter of $\mathbf{w}$ (end of $\mathbf{w}$). Given each pair i and j let $\mathbf{t}_{ij}$ be a choice of the words with $b(\mathbf{t}_{ij}) = i$ and with $e(\mathbf{t}_{ij}) = j$. Such a choice exists because A is irreducible.

First we show that (a) implies (b). We describe the point with a dense orbit, $\mathbf{s}^*$. List all the words of length one with the proper choice of the transition word $\mathbf{t}_{ij}$ between them to make the sequence allowable. Then list all the allowable words of length two with the proper choice of the transition word $\mathbf{t}_{ij}$ between them to make the sequence allowable. Continue by induction listing all the allowable words of length n with the proper choice of the transition word $\mathbf{t}_{ij}$ between them to make the sequence allowable. In this way we construct an infinite allowable sequence which contains all the allowable words of finite length. If $\mathbf{u}$ is a sequence in Σ_A and V is a neighborhood of $\mathbf{u}$, then there is some n such that any sequence which agrees with $\mathbf{u}$ in the first n places is contained in V. Now for this n there is a word somewhere in $\mathbf{s}^*$ which agrees with this word of length n in $\mathbf{u}$. Next there is a k such that $\sigma_A^k(\mathbf{s}^*)$ has this word in the first n places. Thus $\sigma_A^k(\mathbf{s}^*) \in V$. Since $\mathbf{u}$ and V were arbitrary, this proves that the orbit of $\mathbf{s}^*$ is dense. (The reader might consider the case when A is a permutation matrix separately, but the above proof is also applies to this case.)

Next we show that (b) implies (a). If A is a permutation matrix, then it is clearly the case that A is irreducible. Thus we can assume that A is not a permutation matrix. Take $\mathbf{s}^* \in \Sigma_A$ whose orbit is dense in Σ_A. Because the matrix is not a permutation matrix, $\mathbf{s}^*$ can not be a periodic point.

Take an arbitrary pair i and j. By assumption (ii) for a transition matrix, it is possible to take $\mathbf{a} \in \Sigma_A$ such that $a_0 = i$. If a point $\mathbf{t}$ is close enough to $\mathbf{a}$ then $t_0 = a_0$. There is some k_1 such that $\sigma_A^{k_1}(\mathbf{s}^*)$ is within this distance so $\sigma_A^{k_1}(\mathbf{s}^*)_0 = a_0 = i$, and $s^*_{k_1} = a_0 = i$. Thus i appears in the sequence for $\mathbf{s}^*$.

Similarly there is a $\mathbf{b} \in \Sigma_A$ such that $b_0 = j$. By Lemma 2.5, $\sigma_A^{k_1}(\mathbf{s}^*)$ has a dense forward orbit. The same argument as above shows there is a $k_2 > k_1$ with $\sigma_A^{k_2}(\mathbf{s}^*)_0 = b_0 = j$, and so $s^*_{k_2} = b_0 = j$. Thus there is an allowable word in $\mathbf{s}^*$ from the k_1^{th} entry to the k_2^{th} entry which goes from i to j, and we can get from i to j for an arbitrary pair i and j. □

By Lemma 2.4, in order to assume that A is not a permutation matrix it is only necessary to assume that $\sum_j a_{ij} \geq 2$ for some i. We use this assumption to prove that Σ_A is perfect. First, we prove a preliminary lemma.

Lemma 2.7. *Assume that A is an irreducible transition matrix such that $\sum_j a_{i_0 j} \geq 2$ for some i_0. Then for each i there exists a $k = k(i)$ for which $\sum_j (A^k)_{ij} \geq 2$.*

PROOF. Since A is irreducible there is a word $\mathbf{w} \in \Sigma_A$ such that $b(\mathbf{w}) = i$ and $e(\mathbf{w}) = i_0$. Let the length of $\mathbf{w}$ be k, so there are $k-1$ transitions. Thus $\text{num}(k-1, i, i_0) \geq 1$. Then there are least two possible choices after i_0. Thus

$$\sum_j (A^k)_{ij} \geq \sum_j (A^{k-1})_{ii_0} a_{i_0 j} = \sum_j \text{num}(k-1, i, i_0) a_{i_0 j} \geq 2.$$

□

Proposition 2.8. *Assume that A is an irreducible transition matrix with $\sum_j a_{i_0 j} \geq 2$ for some i_0. Then Σ_A is perfect.*

PROOF. For each i there is a $k = k(i)$ such that $\sum_j (A^k)_{ij} \geq 2$. Take an $\mathbf{s} \in \Sigma_A$. Take a cylinder set U as a neighborhood, $U = \{\mathbf{t} \in \Sigma_A : t_i = s_i \text{ for } 0 \leq i \leq n\}$. Then there exists a $k = k(s_n)$ such that $\sum_j (A^k)_{s_n j} \geq 2$. Because there is more than one choice for the transitions from the n^{th} to the $(n+k)^{th}$ entry, there is a $\mathbf{t} \in U$ with $t_{n+m} \neq s_{n+m}$ for some m with $1 \leq m \leq k$. This is true for all $\mathbf{s} \in \Sigma_A$ and all cylinder sets, so Σ_A is perfect. □

Proposition 2.9. *Assume that A is an eventually positive transition matrix. Then σ_A is topologically mixing on Σ_A.*

We leave the proof to the exercises. See Exercise 3.14.

Proposition 2.10. *Assume A is a transition matrix. (We do not assume A is irreducible.) Then the states can be ordered in such a way that A has the following block form:*

$$A = \begin{pmatrix} A_1 & * & * & \cdots & * & * \\ 0 & A_2 & * & \cdots & * & * \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & A_m \end{pmatrix}$$

where (i) each A_j is irreducible, (ii) the $*$ terms are arbitrary, and (iii) all the terms below the blocks A_j are all 0. *Moreover, the nonwandering set $\Omega(\sigma_A) = \Sigma_{A_1} \cup \cdots \cup \Sigma_{A_m}$.*

We defer the proof to the exercises. See Exercise 3.15.

REMARK 2.1. For an introductory treatment of further topics in symbolic dynamics, see Boyle (1993).

3.3 Zeta Function

Artin and Mazur had the idea to combine the number of periodic points of all periods into a single invariant (Artin and Mazur, 1965). If we list all these numbers there are countably many invariants. For certain classes of maps, when these numbers are combined together in a certain way they yield a rational function which has only a finite number of coefficients. Thus the information given by these countable number of invariants is contained in this finite set of coefficients. We proceed with the formal definitions.

Definition. Let $f : X \to X$ be a map, and $N_k(f) = \#(\mathrm{Per}_k(f)) = \#\,\mathrm{Fix}(f^k)$. The *zeta function for* f is defined to be

$$\zeta_f(t) = \exp\Big(\sum_{k=1}^{\infty} \frac{1}{k} N_k(f) t^k\Big).$$

The zeta function is clearly invariant under topological conjugacy because the number of points of each period is preserved. For more discussion of the zeta function see Chapter 5 of Franks (1982). In this section, we merely calculate the zeta function for a subshift of finite type. This theorem was originally proved by Bowen and Lanford (1970). In Chapter VII, we return to prove that the zeta function is a rational function of t for some further types of maps (toral Anosov diffeomorphisms). Before stating the theorem, we give a connection between the determinant, exponential, and trace of a matrix. (The exponential of a matrix is defined by substituting the matrix into the power series for the exponential. It is discussed further in Section 4.3 in the context of solutions of linear differential equations.)

Lemma 3.1 (Liouville's Formula). *Let B be a matrix. Then*

$$\det(e^B) = e^{\mathrm{tr}(B)}.$$

PROOF. Let $\mathbf{e}_1 \ldots \mathbf{e}_n$ be the standard basis. The following calculation uses the facts that the determinant is alternating in the columns, that $e^{Bt} = (e^{Bt}\mathbf{e}_1, \ldots, e^{Bt}\mathbf{e}_n)$, that $e^{B0} = I$, and that $\frac{d}{dt} e^{Bt} = B e^{Bt}$. Then

$$\begin{aligned}
\frac{d}{dt}\det(e^{Bt})|_{t=0} &= \sum_j \det(\mathbf{e}_1, \ldots, \mathbf{e}_{j-1}, \frac{d}{dt} e^{Bt}\mathbf{e}_j|_{t=0}, \mathbf{e}_{j+1}, \ldots, \mathbf{e}_n) \\
&= \sum_j \det(\mathbf{e}_1, \ldots, \mathbf{e}_{j-1}, B\mathbf{e}_j, \mathbf{e}_{j+1}, \ldots, \mathbf{e}_n) \\
&= \sum_j \det(\mathbf{e}_1, \ldots, \mathbf{e}_{j-1}, \sum_i b_{ij}\mathbf{e}_i, \mathbf{e}_{j+1}, \ldots, \mathbf{e}_n) \\
&= \sum_{i,j} b_{ij} \det(\mathbf{e}_1, \ldots, \mathbf{e}_{j-1}, \mathbf{e}_i, \mathbf{e}_{j+1}, \ldots, \mathbf{e}_n) \\
&= \sum_j b_{jj} \det(\mathbf{e}_1, \ldots, \mathbf{e}_{j-1}, \mathbf{e}_j, \mathbf{e}_{j+1}, \ldots, \mathbf{e}_n) \\
&= \mathrm{tr}(B).
\end{aligned}$$

For $t = t_0$,

$$\frac{d}{dt}\det(e^{Bt})|_{t=t_0} = \frac{d}{dt}\det(e^{B(t-t_0)})|_{t=t_0}\det(e^{Bt_0})$$
$$= \mathrm{tr}(B)\det(e^{Bt_0}).$$

Solving the scalar differential equation $\frac{d}{dt}\det(e^{Bt}) = \mathrm{tr}(B)\det(e^{Bt})$ with initial condition $\det(e^{B0}) = 1$ gives that $\det(e^{Bt}) = e^{\mathrm{tr}(B)t}$. Evaluating this solution at $t = 1$ gives the result. □

Theorem 3.2. *Let $\sigma_A : \Sigma_A \to \Sigma_A$ be the subshift of finite type for $A = (a_{ij})$ with $a_{ij} \in \{0, 1\}$ for every pair of i and j. Then the zeta function of σ_A is rational. Moreover $\zeta_{\sigma_A}(t) = [\det(I - tA)]^{-1}$.*

PROOF. By Corollary 3.3, $N_k(\sigma_A) = \mathrm{tr}(A^k)$. Therefore, using the linearity of the trace, the power series expansion of the logarithm, and Lemma 3.10 we can make the following calculation.

$$\begin{aligned}
\zeta_{\sigma_A}(t) &= \exp\big(\sum_{k=1}^{\infty}\frac{1}{k}t^k\,\mathrm{tr}(A^k)\big)\\
&= \exp\big(\mathrm{tr}(\sum_{k=1}^{\infty}\frac{1}{k}t^kA^k)\big)\\
&= \exp\big(\mathrm{tr}(-\log(I - tA))\big)\\
&= \det\big(\exp(\log(I - tA)^{-1})\big)\\
&= \det\big((I - tA)^{-1}\big)\\
&= \big(\det(I - tA)\big)^{-1}.
\end{aligned}$$

This proves the theorem. □

3.4 Period Doubling Cascade

The Sharkovskii Theorem tells us which periods imply which other periods. In particular, if a map $f : \mathbf{R} \to \mathbf{R}$ has finitely many periodic orbits then all the periods must be powers of 2. For the quadratic family, $F_\mu(x) = \mu x(1 - x)$, we saw that it had only fixed points for $0 < \mu \le 3$, and only fixed points and a point of period 2 for $3 < \mu \le 1 + 6^{1/2}$. In fact Douady and Hubbard (1985) proved that for the quadratic family as μ increases new periods are added to the list of periods appearing and never disappear once they have occurred. See de Melo and Van Strien (1993). Let μ_n be the infimum of the parameter values $\mu > 0$ for which F_μ has a point of period 2^n. By the Sharkovskii Theorem, $\mu_n \le \mu_{n+1}$. Notice that all the $\mu_n < 4$ because F_4 has points of all periods. Let μ_∞ be the limiting value of the μ_n as n goes to infinity. The dynamics for F_{μ_∞} is like the map f_∞ given in Exercise 3.8: there is an invariant set on which F_{μ_∞} acts like an adding machine. This sequence of bifurcations is often called the *period doubling route to chaos.*

At the bifurcation value $\mu_1 = 3$ for the family F_μ, the fixed point p_μ changes from attracting for $1 < \mu < \mu_1$ to repelling for $\mu_1 < \mu$. At $\mu = \mu_1$, $F'_{\mu_1}(p_{\mu_1}) = -1$. For μ slightly larger than μ_1, the 2-periodic orbit $\mathcal{O}(p_{\mu,1})$ is attracting with derivative just less than one, $1 > (F^2_\mu)'(p_{\mu,1}) > 0$. In Chapter VI, we study the period doubling bifurcation and show at $\mu = \mu_2$ where the period four orbit is created that $(F^2_{\mu_2})'(p_{\mu_2,1}) = -1$. Again, this 2-periodic orbit $\mathcal{O}(p_{\mu,1})$ changes from attracting to repelling as μ moves past μ_2. The period 4 orbit $\mathcal{O}(p_{\mu,2})$ is initially attracting for μ just slight larger than μ_2 and becomes repelling for $\mu > \mu_3$. This process repeats itself; at $\mu = \mu_n$ the period 2^n orbit $\mathcal{O}(p_{\mu,n})$ is added. This orbit is attracting for $\mu_n < \mu < \mu_{n+1}$ and becomes repelling for $\mu > \mu_{n+1}$.

A natural question to ask is the rate of convergence of the parameter values μ_n to μ_∞. Consider a geometric sequence of numbers, $\lambda_n = C_0 - C_1\lambda^n$, where $0 < \lambda < 1$. For this example the limiting value $\lambda_\infty = C_0$ and λ (or λ^{-1}) gives the rate of convergence to λ_∞. In general, we want to define a quantity which measures the geometric rate of convergence to the limiting value. Feigenbaum (1978) calculated the rate of convergence by means of the limit

$$\delta = \lim_{n\to\infty} \frac{\mu_n - \mu_{n-1}}{\mu_{n+1} - \mu_n}.$$

This value δ is called the *Feigenbaum constant.* Notice for the sequence $\mu_n = C_0 - C_1\lambda^n$, the value δ would equal λ^{-1}:

$$\begin{aligned}\lim_{n\to\infty} \frac{\mu_n - \mu_{n-1}}{\mu_{n+1} - \mu_n} &= \lim_{n\to\infty} \frac{C_0 - C_1\lambda^n - C_0 + C_1\lambda^{n-1}}{C_0 - C_1\lambda^{n+1} - C_0 + C_1\lambda^n} \\ &= \lambda^{-1}.\end{aligned}$$

Feigenbaum (1978) discovered that this constant is the same for several different families of functions. The value has been calculated to be $\delta = 4.669202\ldots$. Both Feigenbaum (1978) and Coullet and Tresser (1978) suggested using the renormalization method to prove the universality of this constant, i.e., that the constant is the same for any one parameter family of functions which go through the period doubling sequence of bifurcations. Much of this program has now been proved by Feigenbaum, Coullet, Tresser, Collet, Eckmann, Lanford, and others, but there are some mathematical aspects of this program which are still unproven. See de Melo and Van Strien (1993), Collet and Eckmann (1980), and Lanford (1984a, 1984b, 1986).

How can these parameter values μ_n be determined for the family F_μ? We mentioned above that the period 2^n orbit is attracting for $\mu_n < \mu < \mu_{n+1}$. In fact the critical point $x_0 = 0.5$ must converge to the attracting periodic orbit. See the discussion of negative Schwarzian derivative in Devaney (1989) or de Melo and Van Strien (1993). Thus to find the attracting periodic orbit we could iterate the critical point a number of times, say 1000, without recording the iterates, and then record or plot the next 1000 iterates. See Figure 4.1A. Note that for the family F_μ the limiting parameter value $\mu_\infty = 3.5699456$. Therefore the whole period doubling bifurcation is shown in Figure 4.1B. Since this part of the orbit is near an attracting periodic orbit, we could inspect the orbit to determine the period. By varying μ, we could determine the value of μ when the orbit changed from period 2^{n-1} to 2^n; this value of μ gives μ_n.

A second method to determine the μ_n is to note that the period 2^{n-1} orbit $\mathcal{O}(p_{\mu,n-1})$ becomes unstable at μ_n and

$$(F^{2^{n-1}}_{\mu_n})'(p_{\mu_n,n-1}) = -1.$$

Thus we could use a numerical scheme (e.g. Newton's method) to search for a point and a parameter value with this property. This search would determine the μ_n.

Finally, there is a third method for determining the rate of convergence given by the Feigenbaum constant by determining slightly different parameter values. We mentioned above that

$$(F^{2^n}_{\mu_n})'(p_{\mu_n,n}) = 1 \qquad \text{and} \qquad (F^{2^n}_{\mu_{n+1}})'(p_{\mu_{n+1},n}) = -1.$$

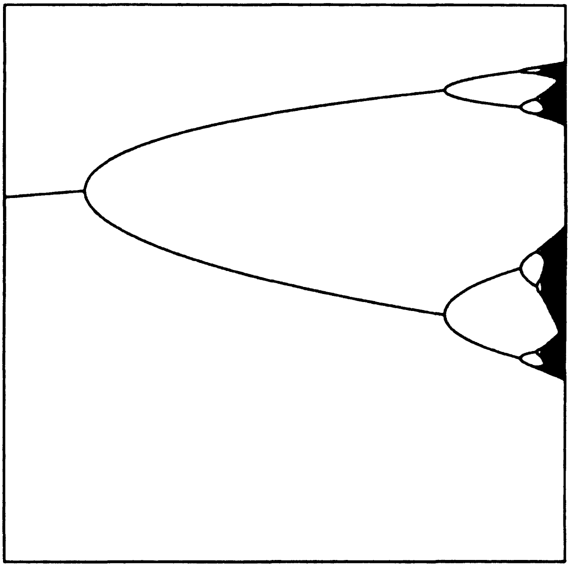

FIGURE 4.1A. The Bifurcation Diagram for the Family F_μ: the Horizontal Direction is the Parameter μ Between 2.9 and 3.6; the Vertical Direction is the Space Variable x Between 0 and 1

Between these two parameter values, there is another value μ'_n for which

$$(F^{2^n}_{\mu'_n})'(p_{\mu'_n,n}) = 0,$$

i.e., the critical point 0.5 has least period 2^n. These parameter values satisfy $\cdots < \mu_n < \mu'_n < \mu_{n+1} < \cdots$. Using these parameter values μ'_n instead of the μ_n gives the same universal constant as the rate of convergence.

For larger values of the parameter μ but with μ still less than 4, an orbit of a point for the quadratic family F_μ seems to be dense in the whole interval $[0, 1]$. In fact Jakobson (1971, 1981) proved that there is a set of parameter values $\mathcal{M} \subset [4 - \epsilon, 4]$ such that (i) $\mathcal{M}$ has positive Lebesgue measure (and 4 is a density point) and (ii) for every $\mu \in \mathcal{M}$, F_μ has an invariant measure ν_μ on $[0, 1]$ that is absolutely continuous with respect to Lebesgue measure on $[0, 1]$. This result implies that most points in $[0, 1]$ have orbits which are dense in the interval $[0, 1]$ for these parameter values. Many people have written papers on this and related results. See de Melo and Van Strien (1993) for further discussion of this result.

Recently, Benedicks and Carleson (1991) have used results about the transitivity of this one dimensional family of maps to prove the transitivity of the two dimensional Hénon family of maps for certain parameter values. We discuss this further in Chapter VII.

3.5 Chaos

Dynamical systems are often said to exhibit chaos without a precise definition of what this means. In this section, we discuss concepts related to the chaotic nature of maps and give some tentative definitions of a chaotic invariant set.

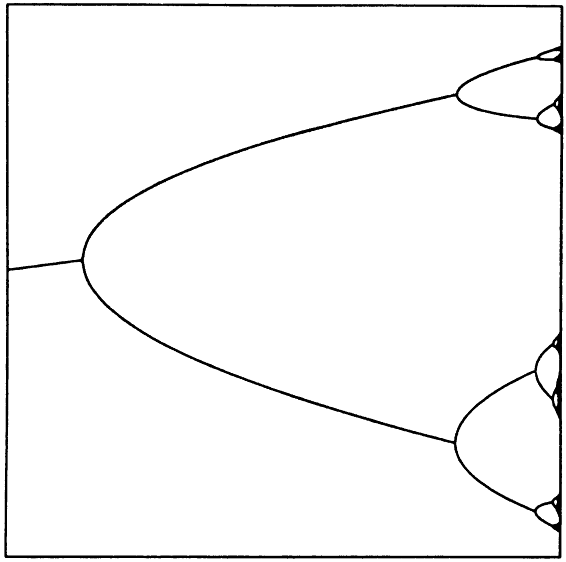

FIGURE 4.1B. The Bifurcation Diagram for the Family F_μ: the Horizontal Direction is the Parameter μ Between 3.54 and 3.5701; the Vertical Direction is the Space Variable x Between 0.47 and 0.57

In Section 2.5, we prove that the quadratic map F_μ, with $\mu > 4$, is transitive on its invariant set Λ_μ. The property of being transitive implies that this set can not be broken up into two closed disjoint invariant sets. For a set to be called "chaotic" it should be (dynamically) indecomposable in some sense of the word. A weaker notion than transitive, which still includes the some kind of dynamic indecomposability of an invariant set, is that it is chain transitive. See Section 2.3. This latter condition seems more natural than topological transitivity but it is not as strong and allows certain examples which do not seem chaotic. (Both periodic motion and "quasi-periodic" motion is chain transitive but not very chaotic.) Therefore in the first definition of a chaotic invariant set we require the map to be topologically transitive and only give chain transitivity as an alternative assumption.

To define a chaotic invariant set, we also want to add a second assumption which indicates that the dynamics of the map on the invariant set are disorderly, or at least that nearby orbits do not stay near each other under iteration. The following definition of sensitive dependence on initial conditions is one possible such concept. In the next section we define the Liapunov exponents of a map. Another way to express that nearby orbits diverge is that the map has positive Liapunov exponents. See Section 3.6.

Definition. A map f on a metric space X is said to have *sensitive dependence on initial conditions* provided there is an $r > 0$ (independent of the point) such that for each point $\mathbf{x} \in X$ and for each $\epsilon > 0$ there is a point $\mathbf{y} \in X$ with $d(\mathbf{x},\mathbf{y}) < \epsilon$ and a $k \geq 0$ such that $d(f^k(\mathbf{x}), f^k(\mathbf{y})) \geq r$.

One of the early situations where sensitive dependence was observed was in a set of differential equations in three variables. E. Lorenz was studying a the system mentioned in Section 1.3 and discussed in Section 7.10 (Lorenz, 1963). While numerically integrating the equations, he recorded the coordinates of the trajectory to only a three-decimal-place accuracy. After calculating an orbit, he tried to duplicate the latter part of the trajectory by entering as a new initial point $\mathbf{q}$ the coordinates of some point part way through the initial calculation, $\mathbf{p}_{t_0}$. Because the original trajectory had more decimal places stored in memory than he entered the second time by hand, the points $\mathbf{p}_{t_0}$ and $\mathbf{q}$ were not the same but merely nearby points. He observed that the two trajectories, the original trajectory and the one started with the slightly different initial condition, followed each other for a period of time and then diverged from each other rapidly. This divergence is an indication of sensitive dependence on initial conditions of the particular system which he was studying.

Another way that sensitive dependence is manifest is through the round off errors of the computer. Curry (1979) reports on numerical studies of the Hénon map (for $A = 1.4$ and $B = -0.3$) using two different computers. After 60 iterates, the iterates have nothing to do with each other. Thinking of the numerical orbit on a computer as an ϵ-chain of the function, two different ϵ-chains diverge, giving an indication of sensitive dependence on initial conditions. On the other hand, the plot of the orbits on the two machines seem to fill up the same subset of the plane, giving an indication that the function is topologically transitive (or at least chain transitive) on this invariant set. For further discussion of an attractor for the Hénon map, see Sections 1.3 and 7.9.

The concept of sensitive dependence on initial conditions is closely related to another concept called expansive: a map is expansive provided any two orbits become at least a fixed distance apart.

Definition. A map f on a metric space X is said to be *expansive* provided there is an $r > 0$ (independent of the points) such that for each pair of points $\mathbf{x}, \mathbf{y} \in X$ there is a $k \geq 0$ such that $d(f^k(\mathbf{x}), f^k(\mathbf{y})) \geq r$. If f is a homeomorphism, then in the definition of expansive we allow $k \in \mathbb{Z}$ and do not require that k is positive, i.e., there is an $r > 0$ such that for each pair of points $\mathbf{x}, \mathbf{y} \in X$ there is a $k \in \mathbb{Z}$ such that $d(f^k(\mathbf{x}), f^k(\mathbf{y})) \geq r$.

If f is expansive and X is a perfect metric space, then it has sensitive dependence on initial conditions. In determining the proper characterization of chaos, the assumption that the map is expansive seems too strong. Therefore we make the following definitions.

Definition. A map f on a metric space X is said to be *chaotic on an invariant set* Y or exhibits *chaos* provided (i) f is transitive on Y and (ii) f has sensitive dependence on initial conditions on Y.

REMARK 5.1. The use of the term chaos was introduced into Dynamical Systems by Li and Yorke (1975). They proved that if a map on the line had a point of period three, then it had points of all periods. They also proved that if a map f on the line has a point of period three, then it has an invariant set S such that

$$\limsup_{n\to\infty} |f^n(p) - f^n(q)| > 0 \qquad \text{and}$$
$$\liminf_{n\to\infty} |f^n(p) - f^n(q)| = 0$$

for every $p, q \in S$ with $p \neq q$. They considered a map with this latter property as chaotic. This property is certainly related to sensitive dependence on initial conditions.

REMARK 5.2. Devaney (1989) gave an explicit definition of a chaotic invariant set in an attempt to clarify the notion of chaos. To our two assumptions, he adds the assumption that the periodic points are dense in Y. Although this property is satisfied by "uniformly hyperbolic" maps like the quadratic map, it does not seem that this condition is at the heart of the idea that the system is chaotic. (This last comment is made even though in the original paper Li and Yorke (1975) proved the existence of periodic points.) Therefore we leave out conditions about periodic points in our definition of chaos.

The paper of Banks, Brooks, Cairns, Davis, and Stacey (1992) proves that any map which (i) is transitive and (ii) has dense periodic points also must have sensitive dependence on initial conditions. However as stated above, we consider the conditions that the map (i) is transitive and (ii) has sensitive dependence on initial conditions a more dynamically reasonable choice of conditions in the definition.

REMARK 5.3. As stated above, an alternative definition of a *chaotic invariant set* Y is that (i) f is chain transitive on Y and (ii) f has sensitive dependence on initial conditions on Y.

This definition allows the following example which does not seem chaotic. Let x and y both be mod 1 variables, so $\{(x, y) : x, y \bmod 1\}$ is the two torus, $\mathbb{T}^2$. Let

$$f(x, y) = (x + y, y)$$

be a shear map. Then f preserves the y variable. The rotation in the x direction depends on the y variable. This map is chain transitive on $\mathbb{T}^2$ but not topologically transitive. The controlled nature of the trajectories make it seem non-chaotic.

One way to avoid the above example and still use chain transitivity as the notion of indecomposability is to require that solutions diverge at an exponential rate. This is defined in the next section in terms of the Liapunov exponents. (Also see Section 8.2 for Liapunov exponents in higher dimensions.) Using this concept, we give an an alternative definition of a chaotic invariant set: an invariant set Y could be called *chaotic* provided that (i) f is chain transitive on Y and (ii) f has a positive Liapunov exponents on Y. Ruelle (1989a) has a long discussion of a chaotic attractor in which he includes the requirements that it be irreducible and has a positive Liapunov exponent.

There is another measurement related to chaos which relates to the invariant set for the system. There are various concepts of (fractal) dimensions including the box dimension which allow the dimension to be an noninteger value. We give some of these dimensions in Section 8.4 (in higher dimensions where the concepts seem to have their natural setting). For experimental data (without specific equations), Liapunov exponents are not very computable but the box dimension of the invariant set is computable. Therefore in the setting of experimental measurements, the box dimension seems like a reasonable measurement of chaos. See the discussion in Chapter 5 of Broer, Dumortier, van Strien, and Takens (1991).

REMARK 5.4. A more mathematical solution to making the notion of chaos precise is in terms of a quantitative measurement of chaos called topological entropy which is defined in Section 8.1. The topological entropy of a map f is denoted by $h(f)$ and is a number greater than or equal to zero and less than or equal to infinity. This quantity has a complicated definition, but can be thought of as a quantitative measurement of the amount of sensitive dependence on initial conditions of the map. If the nonwandering set of f is a finite number of periodic points then $h(f) = 0$. In this case a transitive invariant set is just a single periodic orbit, and this does not have sensitive dependence on initial conditions. If the dynamics of f are complicated as for the quadratic map F_μ on Λ_μ for $\mu > 4$, then $h(F_\mu) > 0$. Therefore, another characterization of a chaotic invariant set Λ for f might be that $h(f|\Lambda) > 0$. Using this definition, we do not need to add the condition that f is transitive: if $h(f|\Lambda) > 0$, then with mild assumptions there is an invariant subset $\Lambda' \subset \Lambda$ on which f is transitive and for which $h(f|\Lambda') > 0$. (This last statement is not obvious but is true based on results of Chapter IX as well as the more precise definition of topological entropy in Section 8.1.)

REMARK 5.5. Thus we have given four alternative definitions of a chaotic invariant set. The characterization of chaos in terms of topological entropy is the most satisfactory one from a mathematical perspective but is not very computable in applications (with a computer). The definition in terms of Liapunov exponents is the most computable (possible to estimate) on a computer. The box dimension is most computable for data from experimental work. Thus there are a number of related important concepts, each of which is important in the appropriate setting. We use the definition of chaos which requires sensitive dependence on initial conditions and topological transitivity for the definition of chaos. The other concepts we refer to by stating the system (i) has positive topological entropy, (ii) has a positive Liapunov exponent, or (iii) has fractional box dimension.

With the above definitions, we can state a result about the quadratic map.

Theorem 5.1. *(a) The shift map σ is chaotic on the full p-shift space, Σ_p. In fact, σ is expansive on Σ_p.*

(b) For $\mu > 4$, the quadratic map F_μ is chaotic on its invariant Cantor set Λ_μ, i.e., $F_\mu|\Lambda_\mu$ has sensitive dependence on initial conditions and is topological transitivity. In fact, $F_\mu|\Lambda_\mu$ is expansive.

PROOF. We proved in an earlier section that both σ and F_μ are transitive on their respective spaces. (The fact that F_μ is transitive follows from the conjugacy to σ.)

To show that σ is expansive and so has sensitive dependence, let $r = 1$. If $\mathbf{s} \neq \mathbf{t}$ for two points in Σ_p, then there is a k such that $s_k \neq t_k$. Then $\sigma^k(\mathbf{s})$ and $\sigma^k(\mathbf{t})$ differ in the 0-th place and $d(\sigma^k(\mathbf{s}), \sigma^k(\mathbf{t})) \geq 1$. This proves that σ is expansive.

For F_μ, since the itinerary map h is a homeomorphism, if $x, y \in \Lambda_\mu$ are distinct points with $\mathbf{s} = h(x)$ and $\mathbf{t} = h(y)$, then there is a k with $s_k \neq t_k$. Therefore $F_\mu^k(x)$ and $F_\mu^k(y)$ are in different intervals I_1 and I_2. Since there is a minimum distance r between these two intervals, $|F_\mu^k(x) - F_\mu^k(y)| \geq r$. This proves that F_μ is expansive. □

REMARK 5.6. The fact that F_μ is expansive on Λ_μ also follows from the following general result that a conjugacy between maps on compact sets preserves expansiveness.

Theorem 5.2. *Let $f : X \to X$ be conjugate to $g : Y \to Y$ where both X and Y are compact. Assume g has sensitive dependence (resp. is expansive) on Y. Then f has sensitive dependence (resp. is expansive) on X.*

PROOF. Let $r > 0$ be the constant for g for either sensitive dependence or expansiveness. Let $h : X \to Y$ be the conjugacy. By compactness, h is uniformly continuous. Therefore given the value $r > 0$ as above, there is a $\delta > 0$ such that if $d(\mathbf{p}, \mathbf{q}) < \delta$ in X then $d(h(\mathbf{p}), h(\mathbf{q})) < r$ in Y. Thus if $d(h(\mathbf{p}), h(\mathbf{q})) \geq r$ in Y then $d(\mathbf{p}, \mathbf{q}) \geq \delta$ in X, or denoting the points differently, if $d(\mathbf{p}, \mathbf{q}) \geq r$ in Y then $d(h^{-1}(\mathbf{p}), (h^{-1}(\mathbf{q})) \geq \delta$ in X.

Now we check the sensitive dependence case. Let $\mathbf{x} \in X$ and $\epsilon > 0$. Then there is an $\epsilon' > 0$ such that if $\mathbf{q} \in Y$ is within ϵ' of $\mathbf{y} = h(\mathbf{x})$ then $\mathbf{p} = h^{-1}(\mathbf{q})$ is within ϵ of $\mathbf{x}$. Take such a $\mathbf{q} \in Y$ that is within ϵ' of $\mathbf{y}$ and $k \geq 0$ as given by the condition of sensitive dependence of g at $\mathbf{y}$. Let $\mathbf{p} = h^{-1}(\mathbf{q})$. Then $d(g^k(\mathbf{y}), g^k(\mathbf{q})) \geq r$, so $d(h^{-1}(g^k(\mathbf{y})), h^{-1}(g^k(\mathbf{q}))) \geq \delta$. But $h^{-1}(g^k(\mathbf{y})) = f^k(h^{-1}(\mathbf{y})) = f^k(\mathbf{x})$ and $h^{-1}(g^k(\mathbf{q})) = f^k(h^{-1}(\mathbf{q})) = f^k(\mathbf{p})$. Therefore, $\mathbf{p}$ is within ϵ of $\mathbf{x}$ and $d(f^k(\mathbf{x}), f^k(\mathbf{q})) \geq \delta$. Thus the δ from the uniform continuity works as the distance by which nearby points of f move apart in the condition of sensitive dependence. The proof for expansiveness is similar. □

3.6 Liapunov Exponents

In discussing chaos, we referred to Liapunov exponents which measure the (infinitesimal) exponential rate at which nearby orbits are moving apart. In this section we give a precise definition and calculate the exponents in a few examples. In Section 8.2 we return to discuss Liapunov exponents in higher dimensions.

Definition. Let $f : \mathbf{R} \to \mathbf{R}$ be a C^1 function. For each point x_0 define the *Liapunov exponent of* x_0, $\lambda(x_0)$, as follows:

$$\begin{aligned}\lambda(x_0) &= \limsup_{n\to\infty} \frac{1}{n} \log(|(f^n)'(x_0)|)\\ &= \limsup_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \log(|f'(x_j)|)\end{aligned}$$

where $x_j = f^j(x_0)$. (The first and second limits are equal by the chain rule.) Note that the right hand side is an average along an orbit (a time average) of the logarithm of the derivative.

The definition of these exponents goes back to the dissertation of Liapunov in 1892, see Liapunov (1907). For a treatment from the point of view of time dependent linear differential equations see Cesari (1959) or Hartman (1964). In higher dimensions, the definition is more complicated than the one given above in one dimension. We discuss this situation in Section 8.2.

Next we give three examples where we can calculate or estimate the Liapunov exponents.

Example 6.1. Let

$$T(x) = \begin{cases} 2x & \text{for } 0 \le x \le 0.5 \\ 2(1-x) & \text{for } 0.5 \le x \le 1. \end{cases}$$

be the tent map. If x_0 is such that $x_j = T^j(x_0) = 0.5$ for some j then $\lambda(x_0)$ is not defined because the derivative is not defined. Such points make up a countable set. For other points $x_0 \in [0,1]$, $|f'(x_j)| = 2$ for all j, so the Liapunov exponent, $\lambda(x_0)$, is $\log(2)$.

Example 6.2. Let $F_\mu(x) = \mu x(1-x)$ for $\mu \ge 2 + 5^{1/2}$. Let Λ_μ be the invariant Cantor set. Then for $x_0 \in \Lambda_\mu$, $\log(|F_\mu'(x_j)|) \ge \lambda_0 > 0$ for some λ_0. Thus the average is larger than λ_0, $\lambda(x_0) \ge \lambda_0$. Thus we may not know an exact value, but it is easy to derive an inequality and know that the exponent is positive.

Before giving the last example, we make some connection between the Liapunov exponent and the space average with respect to an invariant measure. If f has an invariant Borel measure μ with finite total measure and support on a bounded interval, then the Birkhoff Ergodic Theorem (Theorem VII.2.2) says that the limit of the quantity defining $\lambda(x_0)$ actually exists, and is not just a lim sup, for μ-almost all points x_0. In fact, since the measure is a Borel measure and $\log(|f'(x)|)$ is continuous and bounded above, $\lambda(x)$ is a measurable function and

$$\int \lambda(x)\, d\mu(x) = \int \log(|f'(x)|)\, d\mu(x).$$

If f is "ergodic" with respect to μ, then $\lambda(x)$ is constant μ-almost everywhere and

$$\lambda(x) = \frac{1}{|\mu|} \int \log(|f'(x)|)\, d\mu(x) \qquad \mu\text{-almost everywhere,}$$

where $|\mu|$ is the total measure of μ. (See Section 7.2 for the definition of ergodic.) This says that the time average of the logarithm of the derivative is equal to the space average (the integral) of the logarithm of the derivative for μ-almost point. The point to understand from this discussion is that if the map preserves a reasonable measure then the Liapunov exponent is constant almost everywhere.

In higher dimensions, the proof that the appropriate limit exists for almost every x requires a much more complicated ergodic theorem due to Oseledec (1968). See Section 8.2.

Example 6.3. Let $F_4(x) = 4x(1-x)$ be the quadratic map for $\mu = 4$.

If x_0 is such that $x_j = F_4^j(x_0) = 0.5$ for some j, then $\log(|F_4'(x_j)|) = \log(|F_4'(0.5)|) = \log(0) = -\infty$. Therefore $\lambda(x_0) = -\infty$ for these x_0.

If $x_0 = 0$ or 1, then $\lambda(x_0) = \log(|F_4'(0)|) = \log(4) > 0$.

For points $x_0 \in (0,1)$ for which x_j is never equal to 0 or 1 (and so never equals 0.5), we use the conjugacy of F_4 with the tent map T, $h(y) = \sin^2(\pi y/2)$. (This conjugacy is verified in Example II.6.2.) Note that h is differentiable on $[0,1]$ so there is a $K > 0$ such that $|h'(y)| < K$ for $y \in [0,1]$. Also $h'(y) > 0$ in the open interval $(0,1)$, so for any (small) $\delta > 0$ there is a bound $K_\delta > 0$ such that $K_\delta < |h'(y)|$ for $h(y) \in [\delta, 1-\delta]$. For x_0 as above,

$$\begin{aligned}
\lambda(x_0) &= \limsup_{n\to\infty} \frac{1}{n} \log(|(F_4^n)'(x_0)|) \\
&= \limsup_{n\to\infty} \frac{1}{n} \log(|(h \circ T^n \circ h^{-1})'(x_0)|) \\
&= \limsup_{n\to\infty} \frac{1}{n} \big(\log(|h'(y_n)|) + \log(|(T^n)'(y_0)|) + \log(|(h^{-1})'(x_0)|) \big) \\
&\le \limsup_{n\to\infty} \frac{1}{n} \big(\log(K) + n\log(2) + \log(|(h^{-1})'(x_0)|) \big) \\
&= \log(2).
\end{aligned}$$

On the other hand for these x_0, we can pick a sequence of integers n_j going to infinity such that $x_{n_j} \in [\delta, 1-\delta]$. Then letting $y_0 = h^{-1}(x_0)$ and $y_n = T^n(y_0)$,

$$\begin{aligned}
\lambda(x_0) &\ge \limsup_{j\to\infty} \frac{1}{n_j} \log(|(F_4^{n_j})'(x_0)|) \\
&= \limsup_{j\to\infty} \frac{1}{n_j} \big(\log(|h'(y_{n_j})|) + \log(|(T^{n_j})'(y_0)|) + \log(|(h^{-1})'(x_0)|) \big) \\
&\ge \limsup_{j\to\infty} \frac{1}{n_j} \big(\log(K_\delta) + n_j\log(2) + \log(|(h^{-1})'(x_0)|) \big) \\
&= \log(2).
\end{aligned}$$

Therefore $\lambda(x_0) = \log(2)$ for all these points. (Note, there are points which repeatedly come near 0.5 but never hit 0.5 for which the limit of the quantity defining the exponent does not exist but only the lim sup.) In particular, the Liapunov exponent is positive for all points whose orbit never hits 0 or 1 (and so never hits 0.5).

Since T preserves Lebesgue measure, the conjugacy also induces an invariant measure μ for F_4; this measure has density function $\pi^{-1}[x(1-x)]^{-1/2}$. Notice the similarity with the density functions we used to prove that F_μ is transitive for $4 < \mu < 2 + 5^{1/2}$. By the above argument, $\lambda(x) = \log(2)$ for μ-almost all points. Integrating with respect to this density function gives that

$$\begin{aligned}\int_0^1 \log(|F_4'(x)|)\,d\mu(x) &= \int_0^1 \frac{\lambda(x)}{\pi[x(1-x)]^{1/2}}\,dx\\ &= \int_0^1 \frac{\log(2)}{\pi[x(1-x)]^{1/2}}\,dx\\ &= \log(2).\end{aligned}$$

On the other hand

$$\begin{aligned}\int_0^1 \log(|F_4'(x)|)\,d\mu(x) &= \int_0^1 \frac{\log(|F_4'(x)|)}{\pi[x(1-x)]^{1/2}}\,dx\\ &= \int_0^1 \log(|T'(y)|)\,dy\\ &= \log(2).\end{aligned}$$

These are equal as the Birkhoff Ergodic Theorem says they must be.

REMARK 6.1. In the last section we mentioned that topological entropy is a measure of complexity of the dynamics of a map. (The formal definition of entropy is given in the Section 8.1.) Katok (1980) has proved that if a map preserves a non-atomic (continuous) Borel probability measure μ for which μ-almost all initial conditions have non-zero Liapunov exponents, then the topological entropy is positive, so the map is chaotic. Thus a good computational criterion for chaos is whether a function has a positive Liapunov exponent for points in a set of positive measure.

3.7 Exercises

Sharkovskii's Theorem

3.1. Let x be a point of period n for f, $f : \mathbf{R} \to \mathbf{R}$ continuous, $\mathcal{O}(x) = \{x_1, \ldots, x_n\}$ be the orbit of x with $x_1 < x_2 < \cdots < x_n$. Let

$$\mathcal{A} = \{[x_1, x_2], [x_2, x_3], \ldots, [x_{n-1}, x_n]\}$$

be the set of intervals induced by the orbit. Assume there exists $I_1 \in \mathcal{A}$ such that $f(I_1) \supset I_1$. Define $J_1 = I_1$, and define inductively

$$J_j = \bigcup\{L \in \mathcal{A} : f(J_{j-1}) \supset L\}$$

for $j = 2, \ldots, n-1$. Further assume that $J_{n-1} \supset K$ for $K \in \mathcal{A}$, where n is the period of x.

(a) Show that there exists $I_{n-2} \in \mathcal{A}$ such that I_{n-2} f-covers K and $I_{n-2} \subset J_{n-2}$.
(b) Show that there exists a sequence $I_j \in \mathcal{A}$ for $j = 1, \ldots, n-1$ with I_1 as above, $I_{n-1} = K$ and such that I_j f-covers I_{j+1} for $j = 1, \ldots, n-2$.

3.2. Assume that $f : \mathbf{R} \to \mathbf{R}$ is continuous and has a point of period n. Assume $n = 2^m p$ with $p > 1$ odd, $m \geq 1$, and n is maximal in the Sharkovskii ordering. Further assume that $k = 2^s q$ with $s \geq m+1$ and $1 \leq q$ and q odd. Prove that f has a point of period k. Thus prove Case 3a of Sharkovskii's Theorem, page 70.

3.3. Assume that $f : \mathbf{R} \to \mathbf{R}$ is continuous and has a point of period n. Assume $n = 2^m p$ with $p > 1$ odd, $m \geq 1$, and n is maximal in the Sharkovskii ordering. Further assume that $k = 2^s$ with $s \leq m$. Prove that f has a point of period k. Thus prove Case 3b of Sharkovskii's Theorem, page 70.

3.4. Assume that $f : \mathbf{R} \to \mathbf{R}$ is continuous and has a point of period n. Assume $n = 2^m p$ with $p > 1$ odd, $m \geq 1$, and n is maximal in the Sharkovskii ordering. Further assume that $k = 2^m q$ with q odd and $q > p$. Prove that f has a point of period k. Thus prove Case 3c of Sharkovskii's Theorem, page 70.

3.5. Let $n = 2^m p$ with $p > 1$ odd, $m \geq 1$. Prove that there is a continuous function $f : [0,1] \to [0,1]$ whose periods, $\mathcal{P}(f)$, are exactly the set $\{k : n \rhd k\}$.

3.6. Construct a map of $\mathbf{R}$ with points of all periods except 3, i.e., construct a map with all periods implied by period 5 from Sharkovskii's Theorem but no other periods. Hint: Take an orbit of period 5, $\{x_1 < x_2 < x_3 < x_4 < x_5\}$, with the order given as in the proof of Sharkovskii's Theorem; let the map be linear on each interval $[x_i, x_{i+1}]$; show that this map works.

3.7. Construct a map with a point of period 10 and all periods implied by 10 by the Sharkovskii ordering but no others. Hint: Take the double of the map in the problem before the last.

3.8. As defined in Example 1.3, let $f_n : [0,1] \to [0,1]$ be the function with exactly one point of period 2^i for $0 \leq i \leq n$ and no other periodic points. Define f_∞ by $f_\infty(x) = \lim_{n\to\infty} f_n(x)$.

(a) Prove that f_∞ is continuous.
(b) Prove that the periods of f_∞ are exactly $\{2^i : 0 \leq i < \infty\}$.
(c) Prove that for each n, f_∞ has exactly one periodic orbit of each period 2^n, that it is repelling, and that the points of this orbit lie in the gaps $G_{n,j}$ which define the middle-(1/3) Cantor set.
(d) Let

$$S_n = [0,1] \setminus \bigcup_{\substack{1\leq j\leq 2^{k-1}\\ 1\leq k\leq n}} G_{k,j}$$

be the union of the 2^n intervals used to define the middle-(1/3) Cantor set. Prove that $f_\infty(S_n) = S_n$.
(e) Let $\Lambda = \bigcap_{n\geq 1} S_n$. Prove that Λ is invariant for f_∞.
(f) Let Σ_2 be the set of all sequences of 0's and 1's. Define $A : \Sigma_2 \to \Sigma_2$ by

$$A(s_0 s_1 s_2 \ldots) = (s_0 s_1 s_2 \ldots) + (1000\ldots) \bmod 2,$$

i.e., $(1000\ldots)$ is added to $(s_0 s_1 s_2 \ldots)$ mod 2 with carrying (so $(11\bar{0}) + (1\bar{0}) = (001\bar{0})$). The map A on Σ_2 is called the *adding machine*. Define $h : \Lambda \to \Sigma_2$ by $h(p) = \mathbf{s}$ where $s_k = 1$ if p belongs to the left hand choice of the interval in S_{n-1}. Prove that h is a topological conjugacy from f_∞ on Λ to A on Σ_2.
(g) Prove that the adding machine A on Σ_2 has no periodic points, and every forward orbit is dense in Σ.

Subshifts of Finite Type

3.9. Give the matrix of the subshift of finite type for the map in Exercise 3.6 and the intervals $[x_i, x_{i+1}]$.

3.10. Let

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

and

$$A^n = \begin{pmatrix} a_n & b_n \\ c_n & d_n \end{pmatrix}.$$

(a) For a vector (x_0, y_0), let $(x_n, y_n) = (x_0, y_0)A^n$. With $y_{-1} = x_0$, prove that $x_{n+1} = y_n$, and $y_{n+1} = y_n + y_{n-1}$. (This is a Fibonacci sequence.)

(b) Use the fact that $(1,0)A^n = (a_n, b_n)$ and $(0,1)A^n = (c_n, d_n)$ to prove that $a_n = a_{n-1} + a_{n-2}$ and $d_n = d_{n-1} + d_{n-2}$.

(c) Prove that $\mathrm{tr}(A^n) = \mathrm{tr}(A^{n-1}) + \mathrm{tr}(A^{n-2})$.

3.11. Consider the matrix A given in the last problem. Find all the fixed points of σ_A, σ_A^2, σ_A^3, and σ_A^4. Group the points into orbits and give their least period.

3.12. Let A be an n by n matrix with $a_{ij} \in \{0,1\}$, $\sum_j a_{ij} \geq 1$ for all i, and $\sum_i a_{ij} \geq 1$ for all j. We define i to be equivalent to j, $i \sim j$, if there exist $k = k(i,j) \geq 0$ and $m = m(i,j) \geq 0$ such that $(A^k)_{ij} \neq 0$ and $(A^m)_{ji} \neq 0$. (Because we allow $k = 0 = m$, i is equivalent to itself.) Break $\{1, \ldots, n\}$ into equivalence classes, $\{1, \ldots, n\} = S_1 \cup \cdots \cup S_p$ with $S_i \cap S_j = \emptyset$ for $i \neq j$. Assume that for each equivalence class, S_q, there exists a $i_q \in S_q$ such that $\sum_{j \in S_q} a_{i_q j} \geq 2$. Prove that Σ_A is perfect.

3.13. Let $f : \mathbf{R} \to \mathbf{R}$ be a C^1 function. Assume there are p closed and bounded intervals $I_1, I_2, \ldots, I_p$ and $\lambda > 1$ such that (i) $|f'(x)| \geq \lambda$ for all $x \in \bigcup_{i=1}^p I_i \equiv \mathcal{I}$ and (ii) if $f(I_i) \cap I_j \neq \emptyset$ then $f(I_i) \supset I_j$. Let A be the matrix of the subshift of finite type defined by $a_{ij} = 1$ if $f(I_i) \supset I_j$ and $a_{ij} = 0$ if $f(I_i) \cap I_j = \emptyset$. Further assume that (iii) A is transitive and irreducible. Let $\Lambda = \bigcup_{i=1}^p f^{-i}(\mathcal{I})$. Prove that $f|\Lambda$ is conjugate to the subshift of finite type σ_A on Σ_A.

3.14. (This exercise asks you to prove Proposition 2.9, page 77.) Let A be an eventually positive transition matrix, with $(A^k)_{ij} \neq 0$ for all i and j.

(a) Prove for $n \geq k$ that $(A^n)_{ij} \neq 0$ for all i and j.

(b) Prove that σ_A is topologically mixing on Σ_A.

3.15. (This exercise asks you to prove Proposition 2.10, page 77.) Assume A is a transition matrix. (We do not assume A is irreducible.)

(a) Prove that the states can be ordered in such a way that A has the following block form:

$$A = \begin{pmatrix} A_1 & * & * & \cdots & * & * \\ 0 & A_2 & * & \cdots & * & * \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & A_m \end{pmatrix}$$

where (i) each A_j is irreducible, (ii) the $*$ terms are arbitrary, and (iii) all the terms below the blocks A_j are all 0. Hint: Define an ordering on the states as follows. Call $i \geq j$ provided there is a $k = k(i,j)$ such that $A^k_{i,j} \neq 0$. Call states i and j equivalent provided $i \geq j$ and $j \geq i$. Group together the equivalent state and order all the states in terms of the above ordering.

(b) Prove that the nonwandering set $\Omega(\sigma_A) = \Sigma_{A_1} \cup \cdots \cup \Sigma_{A_m}$. Hint: Show that $\Omega(\sigma_{A_j}) = \Sigma_{A_j}$ for all j so $\Omega(\sigma_A) \supset \Sigma_{A_1} \cup \cdots \cup \Sigma_{A_m}$. Also show that all points in $\Sigma_A \setminus (\Sigma_{A_1} \cup \cdots \cup \Sigma_{A_m})$ are wandering.

3.16. Let Σ_A be a subshift of finite type with metric d defined in Chapter II. Let $\delta = 0.5$. Assume $\{\mathbf{s}^{(j)} \in \Sigma_A\}$ is a 0.5-chain for σ_A on Σ_A. Explicitly indicated the point $\mathbf{t} \in \Sigma_A$ which 0.5-shadows this 0.5-chain.

Zeta Functions

3.17. Let A and B each be square matrices and

$$\zeta(t) = \exp\Big(\sum_{k=1}^{\infty} \frac{\mathrm{tr}(A^k) - \mathrm{tr}(B^k)}{k} t^k\Big).$$

Prove that

$$\zeta(t) = \frac{\det(I - tB)}{\det(I - tA)}.$$

Chaos and Liapunov Exponents

3.18. Prove that F_μ is expansive on $\mathbf{R}$ for $\mu > 4$.

3.19. Consider $f_\mu(x) = \mu x \bmod 1$, for $\mu > 1$.

(a) Calculate the Liapunov exponent.

(b) Prove that $f_\mu(x)$ has sensitive dependence on initial conditions.

3.20. Let $f : \mathbf{R} \to \mathbf{R}$ be C^1. Assume p is a periodic point and $\omega(x_0) = \mathcal{O}(p)$. Prove that the Liapunov exponents of x_0 and p are equal, $\lambda(x_0) = \lambda(p)$.

3.21. Let $F_\mu(x) = \mu x(1 - x)$ as usual. For $1 < \mu < 1 + 6^{1/2}$, find the Liapunov exponents for the different points $x \in [0, 1]$. Hint: For $3 < \mu < 1 + 6^{1/2}$, there is an attracting orbit of period 2. See Exercise 2.6. Also see Exercise 3.20.

3.22. Let $0 < \alpha < 0.5$, and $f_\alpha : [0, 1] \to [0, 1]$ be defined by

$$f_\alpha = \begin{cases} 2(\alpha - x) & \text{for } 0 \le x \le \alpha \\ 2(x - \alpha) & \text{for } \alpha \le x \le 0.5 + \alpha \\ 2(0.5 + \alpha - x) + 1 & \text{for } 0.5 + \alpha \le x \le 1. \end{cases}$$

(a) Draw the graph of f_α.

(b) Find the intervals on which f_α is transitive and describe its dynamics.