From: Hasselblat & Katok: A first course in Dynamics, 2003

CHAPTER 4

Recurrence and Equidistribution on the Circle

So far we concentrated on dynamical systems where the asymptotic behavior can be described simply: Every orbit was either fixed (sometimes periodic) or was attracted to (possibly different) fixed points as the time approached positive and negative infinity. In several situations, such as Proposition 2.3.5, we showed that no other behavior is possible.

In this chapter we study a fundamentally different type of behavior. Analysts use the rather innocuous term "quasiperiodic" to describe it and to signify that it is not much more than a generalization of periodic behavior. But from the dynamical point of view this is a starting point for the understanding of *nontrivial recurrence*, the central paradigm of the theory of dynamical systems.

We begin with a careful study of this phenomenon in the simplest possible situation, circle rotations. In the second section this already gives a remarkable array of interesting applications. The final section extends some of our insights to nonlinear circle maps.

4.1 ROTATIONS OF THE CIRCLE

The description of our first example is surprisingly simple; it is, in fact, closely related to some of the linear dynamical systems that appeared in Chapter 3, specifically Section 3.1.8.4 with $\rho = 1$: For a linear system with a pair of complex conjugate eigenvalues of absolute value 1, complex behavior may appear on the invariant circles $r = \text{const}$. We now study these rotations of a circle.

4.1.1 Circle Rotations

In Section 2.6.2 we saw two different convenient ways to represent the circle that allow us to write various formulas in a nice fashion. One can use either multiplicative notation, in which the circle is represented as the unit circle in the complex plane

$$S^1 = \{z \in \mathbb{C} \mid |z| = 1\} = \{e^{2\pi i\phi} \mid \phi \in \mathbb{R}\},$$

or additive notation, where

$$S^1 = \mathbb{R}/\mathbb{Z}$$

consists of the real numbers with integer translates identified (recall Figure 2.6.2). In multiplicative notation all algebraic operations make sense as operations over complex numbers. In additive notation we can use addition and subtraction (but not multiplication or division) just as the usual operations over real numbers, but we have to keep in mind that all equalities make sense up to an integer. It is customary to add "(mod 1)" to such equalities. Thus, the expression $a = b \pmod 1$, where a and b are real numbers, means that $a - b$ is an integer.

The logarithm map

$$e^{2\pi i\phi} \mapsto \phi$$

establishes an isomorphism between these representations. Let us measure the length of arcs on the circle by the parameter ϕ; that is, the length of the whole circle is equal to one. Let $\ell(\Delta)$ denote the length of the arc Δ measured in such a way. To similarly define a distance introduce a metric on the set $X = \mathbb{R}/\mathbb{Z} := \{[x] \mid x \in \mathbb{R}\}$ of equivalence classes by setting $d(x, y) := \min\{|b - a| \mid a \in x, b \in y\}$ as in Proposition 2.6.7.

We use the symbol R_α to denote the rotation by the angle $2\pi\alpha$. In multiplicative notation

$$R_\alpha(z) = z_0 z \text{ with } z_0 = e^{2\pi i\alpha}.$$

Not surprisingly, in additive notation we have

$$R_\alpha(x) = x + \alpha \pmod 1. \tag{4.1.1}$$

The iterates of the rotation are correspondingly

$$R_\alpha^n(z) = R_{n\alpha}(z) = z_0^n z$$

in multiplicative notation and

$$R_\alpha^n(x) = x + n\alpha \pmod 1.$$

in additive notation.

A crucial distinction in the dynamics of rotations appears between the cases of the rotation parameter α being rational and irrational.

In the former case, write $\alpha = p/q$, where p, q are relatively prime integers. Then $R_\alpha^q(x) = x$ for all x, so R_α^q is the identity map and after q iterates the transformation simply repeats itself. Thus the total orbit of any point is a finite set and all orbits are q-periodic.

4.1.2 Density of Orbits

The case of irrational α is much more interesting. First, it is clear from the above formulas for the iterates that the orbit of every point is an infinite set. We can, however, say much more.

Proposition 4.1.1 *If $\alpha \notin \mathbb{Q}$, then every positive semiorbit of R_α is dense.*

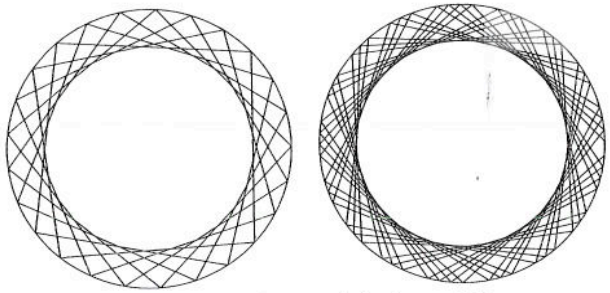


Figure 4.1.1. Periodic orbit and segment of a dense orbit.

Proof Suppose $x, z \in S^1$. To show that z is in the closure of the positive semiorbit of x, let $\epsilon > 0$. The positive semiorbit of x is infinite and no set of $k \geq \lfloor 1/\epsilon \rfloor + 1$ points has pairwise distances all exceeding ϵ. Thus there are $l, k \in \mathbb{N}$ such that $l < k \leq \lfloor 1/\epsilon \rfloor$ and $d(R_\alpha^k(x), R_\alpha^l(x)) < \epsilon$. Then $d(R_\alpha^{k-l}(x), x) < \epsilon$ because R_α^{-l} preserves distances. By the way, this latter distance is independent of x because, if $y \in S^1$, then $y = R_{y-x}(x)$ and

$$\begin{aligned} d\big(R_\alpha^{k-l}(y), y\big) &= d\big(R_\alpha^{k-l}(R_{y-x}(x)), R_{y-x}(x)\big) = d\big(R_{(k-l)\alpha+y-x}(x)\big), R_{y-x}(x)\big) \\ &= d\big(R_{y-x}\big(R_\alpha^{k-l}(x)\big), R_{y-x}(x)\big) = d\big(R_\alpha^{k-l}(x), x\big); \end{aligned}$$

so k and l can be chosen independently of x.

Take $\theta \in [-1/2, 1/2]$ such that $\theta = (k-l)\alpha \pmod 1$. Then $\rho := |\theta| < \epsilon$ and $R_\alpha^{k-l} = R_\theta$. Let $N = \lfloor 1/\rho \rfloor + 1$ (independently of x). Then the subset $\{R_{i\theta}(x) \mid i = 0, 1, \ldots, N\}$ of the positive semiorbit of x divides the circle into intervals of length less than $\rho < \epsilon$, so there is an $n \leq N(k-l)$ such that $d(R_\alpha^n(x), z) < \epsilon$. □

Remark 4.1.2 Since the negative semiorbit of R_α is the positive semiorbit of $R_{-\alpha}$, we also proved the density of negative semiorbits.

An alternate proof of minimality shows the absence of proper invariant closed subsets by contradiction:

Alternate proof of Proposition 4.1.1 Let $A \subset S^1$ be an invariant closed set. The complement $S^1 \setminus A$ is a nonempty open invariant set that consists of disjoint intervals. Let I be the longest of those intervals (or one of the longest, if there are several of the same length). Since rotation preserves the length of any interval, the iterates $R_\alpha^n(I)$ do not overlap. Otherwise, $S^1 \setminus A$ would contain an interval longer than I. Since α is irrational, no iterates of I can coincide, because then an endpoint x of an iterate of I would come back to itself and we would have $x + k\alpha = x \pmod 1$ with $k\alpha = l$ an integer and $\alpha = l/k$ a rational number. Thus the intervals $R_\alpha^n(I)$ are all of equal length and all disjoint, but this is impossible because the circle has finite length and the sum of lengths of disjoint intervals cannot exceed the length of the circle. □

Proposition 4.1.1 motivates the following general definitions.

Definition 4.1.3 A homeomorphism (see Definition A.1.16) $f\colon X \to X$ is said to be *topologically transitive* if there exists a point $x \in X$ such that its orbit $\mathcal{O}_f(x) := (f^n(x))_{n\in\mathbb{Z}}$ is dense in X. Equivalently, every f-invariant open invariant set is dense. A noninvertible map f is said to be topologically transitive if there exists a point $x \in X$ such that its (positive) orbit $\mathcal{O}_f^+(x) := (f^n(x))_{n\in\mathbb{N}_0}$ is dense in X.

The definitions for continuous-time systems are similar.

Definition 4.1.4 A homeomorphism $f\colon X \to X$ is said to be *minimal* if the orbit of every point $x \in X$ is dense in X or, equivalently, if f has no proper closed invariant sets. A closed invariant set is said to be minimal if it contains no proper closed invariant subsets or, equivalently, if it is the orbit closure of any of its points.

Thus Proposition 4.1.1 establishes that any rotation of the circle by an angle incommensurable with π, that is, by an irrational number of degrees (we shall call such a rotation simply an *irrational rotation*), is minimal and hence topologically transitive.

While minimality always implies topological transitivity, the converse is by no means true. Chapter 7 contains various examples that combine topological transitivity (existence of some dense orbits) with the existence of many orbits of different types, for example, infinitely many periodic (finite) orbits whose union is in turn dense.

4.1.3 Dense Orbits

It may be interesting to get a good picture of how an orbit fills the circle densely. We do this in a specific example by following the orbit of 0 under a rotation R_α, where we take

$$\alpha = \cfrac{1}{3 + \cfrac{1}{5 + \cfrac{1}{c}}}$$

for some $c > 1$. $\alpha \in \mathbb{Q}$ if and only if $c \in \mathbb{Q}$. The unusual form of α will seem more natural at the end.

Since $1/4 < \alpha < 1/3$ and hence $3\alpha < 1 < 4\alpha$, the first time the orbit returns more closely to 0 than ever before is after three steps. The first three points, α, 2α, and 3α, are evenly spaced, and since $4\alpha > 1$, 3α is closer to an integer than the previous points. The precise distance is

$$\delta := 1 - 3\alpha = 1 - \cfrac{3}{3 + \cfrac{1}{5 + \cfrac{1}{c}}} = \cfrac{\cfrac{1}{5 + \cfrac{1}{c}}}{3 + \cfrac{1}{5 + \cfrac{1}{c}}} = \cfrac{1}{16 + \cfrac{3}{c}}.$$

To find the next time of closest return we start from the fourth step, using $4\alpha = \alpha - \delta \pmod 1$. So three α-steps take us from α to $\alpha - \delta$. How many of these 3α-steps does it take to get the next closest approach? As before, it should be about α/δ, and the desired number n must satisfy $n\delta < \alpha < (n+1)\delta$. Indeed, $n = 5$ works:

$$5\delta = \frac{5}{15 + \left(1 + \frac{3}{c}\right)} = \frac{1}{3 + \left(\frac{1}{5} + \frac{3}{5c}\right)} < \frac{1}{3 + \frac{1}{5}} < \frac{1}{3 + \frac{1}{5 + \frac{1}{c}}} = \alpha,$$

and

$$6\delta = \frac{6}{16 + \frac{3}{c}} > \frac{6}{18} = \frac{1}{3} > \alpha.$$

These five 3α-steps evenly fill the interval $(0, \alpha)$ and simultaneously its three image intervals. When this next closest return is reached, the orbit segment is a δ-dense subset of the circle spaced evenly (except for the smaller interval of the new closest return). The next closest return after this is determined by c, and it is safe to guess that it will happen after about c steps.

If c were about a billion, this would mean that it takes about a billion 5δ-steps until the next closest return, which is some 15 billion iterations of R_α. In particular, the first 7 billion iterations are guaranteed to leave gaps of $\delta/2 > 1/35$. So large entries in this *continued fraction* form of α are not a good thing for filling the circle well. Continued fractions are discussed in greater detail in Section 15.2.

In conclusion, there is a natural sequence of ever longer time scales during each of which the orbit achieves a finer degree of density in a fairly homogeneous way. Thus, the behavior of an orbit is periodic, except for a little error δ, which produces a perturbation with much higher period – up to an even smaller error, and so on.

4.1.4 Uniform Distribution for Intervals

The preceding discussion suggests that we look into the way orbits of an irrational rotation are distributed on the circle in a quantitative fashion by finding the *frequencies* with which iterates of a point visit various parts of a circle. To be specific, fix an arc $\Delta \subset S^1$, and for $x \in S^1$ and $n \in \mathbb{N}$ let

$$F_\Delta(x, n) := \operatorname{card}\left\{k \in \mathbb{Z} \mid 0 \le k < n,\ R_\alpha^k(x) \in \Delta\right\}.$$

This function is nondecreasing in n for fixed x and Δ. Since the positive semiorbit of any point is dense, there are arbitrarily large positive iterates of x that belong to Δ. Hence

$$F_\Delta(x, n) \to \infty, \text{ as } n \to \infty.$$

The natural measure of how often these visits happen is the *relative frequency* of visits:

$$\frac{F_\Delta(x, n)}{n}. \tag{4.1.2}$$

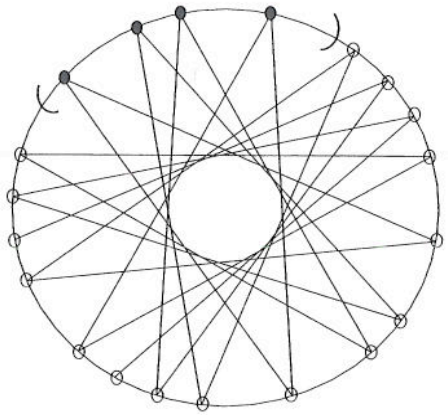


Figure 4.1.2. Frequencies.

Recall that $\ell(\Delta)$ denotes the length of the arc Δ measured by the parameter ϕ introduced at the beginning of Section 4.1.1.

The argument from the proof of Proposition 4.1.1 gives

Proposition 4.1.5 *Suppose α is irrational and consider the rotation R_α. Let Δ, Δ' be arcs such that $\ell(\Delta) < \ell(\Delta')$. Then there exists an $N_0 \in \mathbb{N}$ such that, if $x \in S^1$, $N \ge N_0$, and $n \in \mathbb{N}$, then*

$$F_{\Delta'}(x, n+N) \ge F_\Delta(x, n).$$

Proof By the density of the positive semiorbit of the left end of the arc Δ we can find an $N_0 \in \mathbb{N}$ such that $R_\alpha^{N_0}(\Delta) \subset \Delta'$. Then $R_\alpha^n(x) \in \Delta$ implies $R_\alpha^{n+N_0}(x) \in \Delta'$ and $F_{\Delta'}(x, n+N) \ge F_{\Delta'}(x, n+N_0) \ge F_\Delta(x, n)$ for $N \ge N_0$. □

So far we have not specified what kinds of arcs we consider: open, closed, or half-open. There is no difference as far as limit behavior of the frequencies is concerned, since the difference between the number of visits for an open arc and its closure is at most two. So it is convenient to always take arcs closed on the left and open on the right. For such arcs we have the following *additivity* property: If the right end of Δ_1 coincides with the left end of Δ_2, then $\Delta_1 \cap \Delta_2 = \varnothing$, $\Delta_1 \cup \Delta_2$ is an arc and

$$F_{\Delta_1}(x, n) + F_{\Delta_2}(x, n) = F_{\Delta_1 \cup \Delta_2}(x, n).$$

It is also convenient to define $F_A(x, n) := \operatorname{card}\{k \in \mathbb{Z} \mid 0 \le k < n,\ R_\alpha^k(x) \in A\}$ for any set A that is a union of disjoint arcs. So far we do not know that the limits of relative frequencies exist. However, one can consider the upper limits:

$$\bar{f}_x(A) := \limsup_{n \to \infty} \frac{F_A(x, n)}{n}.$$

These quantities are obviously *subadditive*:

$$\bar{f}_x(A_1 \cup A_2) \le \bar{f}_x(A_1) + \bar{f}_x(A_2).$$

In particular, if $\bigcup_{i=1}^n A_i = S^1$, then $\sum_{i=1}^n \bar{f}_x(A_i) \ge 1$. Proposition 4.1.5 implies

Corollary 4.1.6 *If $\ell(\Delta) < \ell(\Delta')$, then $\bar{f}_x(\Delta) \le \bar{f}_x(\Delta')$.*

Similarly we introduce the lower asymptotic frequencies:

$$\underline{f}_x(A) := \liminf_{n\to\infty} \frac{F_A(x,n)}{n}.$$

Obviously, for any set A we have $F_A(x,n) = n - F_{A^c}(x,n)$, where A^c denotes the complement $S^1 \setminus A$ of A and hence

$$\bar{f}_x(A) = \limsup_{n\to\infty} \frac{F_A(x,n)}{n} = 1 - \liminf_{n\to\infty} \frac{F_{A^c}(x,n)}{n} = 1 - \underline{f}_x(A^c). \tag{4.1.3}$$

Now we can formulate our main statement about asymptotic frequencies:

Proposition 4.1.7 *For any arc $\Delta \subset S^1$ and any $x \in S^1$*

$$f(\Delta) := \lim_{n\to\infty} \frac{F_\Delta(x,n)}{n} = \ell(\Delta),$$

and the limit is uniform in x.

Remark 4.1.8 The property of the sequence $a_n := R_\alpha^n(x)$, $n = 0, 1, 2, \ldots$ expressed by this proposition is called *uniform distribution* or *equidistribution*: The asymptotic frequency of visits is the same for arcs of equal length, regardless of where on the circle they are.

Proof First we show that the frequency of visits cannot be too high.

Lemma 4.1.9 *If $\ell(\Delta) = 1/k$, then $\bar{f}_x(\Delta) \le 1/(k-1)$.*

Proof Consider $k-1$ disjoint arcs $\Delta_1, \Delta_2, \ldots, \Delta_{k-1}$ of length $1/(k-1)$ each. For $1 \le i < k$, Proposition 4.1.5 gives natural numbers N_i such that, if $x \in S^1$, then

$$F_{\Delta_i}(x, n+N_i) \ge F_\Delta(x,n);$$

hence $F_{\Delta_i}(x, n+N) \ge F_\Delta(x,n)$, where $N = \max_i N_i$ and

$$(k-1)F_\Delta(x,n) \le \sum_{i=1}^{k-1} F_{\Delta_i}(x, n+N).$$

Since N is fixed, we let $n \to \infty$ to obtain

$$(k-1)\bar{f}_x(\Delta) \le \bar{f}_x\left(\bigcup_{i=1}^{k-1} \Delta_i\right) = 1. \ \square$$

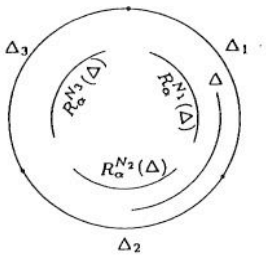


Figure 4.1.3. Upper asymptotic frequencies.

For an arc Δ and $\epsilon > 0$ find k and an arc $\Delta' \supset \Delta$ of length $l/k < \ell(\Delta) + \epsilon$. Then

$$\bar{f}_x(\Delta) < \bar{f}_x(\Delta') < \frac{l}{k-1} < (\ell(\Delta)+\epsilon)\frac{k}{k-1}$$

by Lemma 4.1.9. Letting $\epsilon \to 0$ and thus $k \to \infty$ gives $\bar{f}_x(\Delta) \le \ell(\Delta)$. Combined with (4.1.3) for $A = \Delta^c$, this also gives $\underline{f}_x(\Delta) \ge \ell(\Delta)$. This proves that the limit exists and equals $\ell(\Delta)$. $\square$

4.1.5 Uniform Distribution for Functions

Clearly, frequencies also can be defined for any set A that is a finite union of arcs. To do this in a suggestive way we call

$$\chi_A(x) := \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

the *characteristic function* of A. Then we define

$$F_A(x,n) := \sum_{k=0}^{n-1} \chi_A\left(R_\alpha^k(x)\right),$$

and accordingly the relative frequency is $\sum_{k=0}^{n-1} \chi_A(R_\alpha^k(x))/n$. Since, by definition of the integral, $\ell(\Delta) = \int_{S^1} \chi_\Delta(\phi)d\phi$, Proposition 4.1.7 can be reformulated as

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \chi_A\left(R_\alpha^k(x)\right) = \int_{S^1} \chi_\Delta(\phi)d\phi. \tag{4.1.4}$$

1. Birkhoff Averaging. We can also consider similar expressions for functions φ other than characteristic functions.

Definition 4.1.10 The *Birkhoff averaging operator* $\mathcal{B}_n$ is the operator that associates to a function φ the function $\mathcal{B}_n(\varphi) := \sum_{k=0}^{n-1} \varphi \circ R_\alpha^k / n$ given by

$$\mathcal{B}_n(\varphi)(x) = \frac{1}{n} \sum_{k=0}^{n-1} \varphi\left(R_\alpha^k(x)\right). \tag{4.1.5}$$

Remark 4.1.11 Some useful properties of $\mathcal{B}_n$ are

(1) $\mathcal{B}_n$ is linear: $\mathcal{B}_n(a\varphi + b\psi) = a\mathcal{B}_n(\varphi) + b\mathcal{B}(\psi)$.
(2) $\mathcal{B}_n$ is nonnegative: If $\varphi \ge 0$, then $\mathcal{B}_n(\varphi) \ge 0$. Also, $\mathcal{B}_n$ is positive (or monotone): If $\varphi > 0$, then $\mathcal{B}(\varphi) > 0$.
(3) $\mathcal{B}_n$ is nonexpanding: $\sup_{x\in S^1} \mathcal{B}_n(\varphi)(x) \le \sup_{x\in S^1} \varphi(x)$.
(4) $\mathcal{B}_n$ preserves the average: $\int_{S^1} \mathcal{B}_n(\varphi)(\phi)\, d\phi = \int_{S^1} \varphi(\phi)\, d\phi$.

This leads to the following conclusions:

Proposition 4.1.12

(1) *For any step function φ that is a linear combination of characteristic functions of arcs, $\lim_{n\to\infty} \mathcal{B}_n(\varphi) = \int_{S^1} \varphi(\phi)\, d\phi$.*

(2) *For any function φ that is a uniform limit of step functions we also have* $\lim_{n\to\infty} \mathcal{B}_n(\varphi) = \int_{S^1} \varphi(\phi)\, d\phi$.

Proof Since the map associating to an integrable function its integral over S^1 has properties analogous to those in the remark, we can start from (4.1.4), pass to linear combinations and uniform limits, and compare results.

For the second claim fix $\epsilon > 0$, take a step function φ_ϵ with $\sup_{\phi\in S^1} |\varphi(\phi) - \varphi_\epsilon(\phi)| < \epsilon$, and apply the operators $\mathcal{B}_n$ to $\varphi = \varphi_\epsilon + (\varphi - \varphi_\epsilon)$ to get

$$\begin{aligned}(4.1.6)\quad &\int_{S^1} \varphi(\phi)\, d\phi - 2\epsilon \le \int_{S^1} \varphi(\phi) - \epsilon\, d\phi - \epsilon \le \int_{S^1} \varphi_\epsilon(\phi)\, d\phi - \epsilon \\ &= \lim_{n\to\infty} \mathcal{B}_n(\varphi_\epsilon) - \epsilon \le \liminf_{n\to\infty} \mathcal{B}_n(\varphi) \le \limsup_{n\to\infty} \mathcal{B}_n(\varphi) \le \lim_{n\to\infty} \mathcal{B}_n(\varphi_\epsilon) + \epsilon \\ &= \int_{S^1} \varphi_\epsilon(\phi)\, d\phi + \epsilon \le \int_{S^1} \varphi(\phi) + \epsilon\, d\phi + \epsilon \le \int_{S^1} \varphi(\phi)\, d\phi + 2\epsilon\end{aligned}$$

for any $\epsilon > 0$. □

Lemma 4.1.13 *Every continuous function is the uniform limit of step functions, as is every function with finitely many discontinuity points and with one-sided limits at these points (piecewise continuous functions).*

Proof Every continuous function on S^1 is uniformly continuous; that is, for every $\epsilon > 0$ one can find an $n \in \mathbb{N}$ such that, on every arc of length $1/n$, the function varies by less than ϵ. Dividing S^1 into n such arcs gives a step function that is constant on each arc and differs from the given function by less than ϵ. Essentially the same argument applies to functions with finitely many discontinuity points and one-sided limits at these points. □

The last two results give:

Proposition 4.1.14 *If α is irrational and φ is continuous, then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi\big(R_\alpha^k(x)\big) = \int_{S^1} \varphi(\phi)\, d\phi$$

uniformly in x.

There is a more general class of functions for which the Birkhoff average converges to the integral, namely, all functions integrable in the usual (Riemann) sense.

Theorem 4.1.15 *If α is irrational and φ is Riemann integrable, then*

$$(4.1.7)\qquad \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi\big(R_\alpha^k(x)\big) = \int_{S^1} \varphi(\phi)\, d\phi$$

uniformly in x.

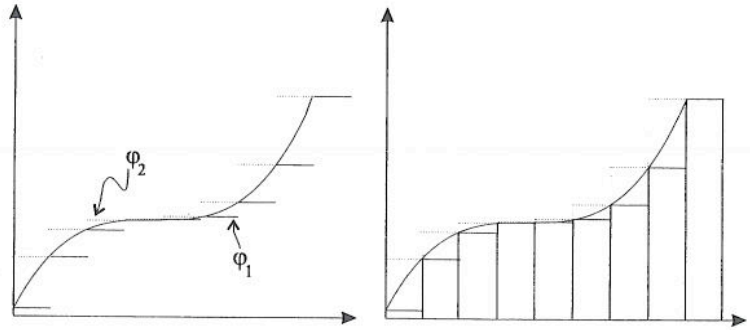


Figure 4.1.4. Approximation by step functions, Riemann sums.

Proof Pick a partition of S^1 into a finite number of arcs I_i. The corresponding lower and upper Riemann sums $\sum_i \min \varphi\restriction_{I_i} l(I_i)$ and $\sum_i \max \varphi\restriction_{I_i} l(I_i)$ can be interpreted as integrals of step functions φ_1 and φ_2 defined by $\varphi_1 = \min \varphi\restriction_{I_i}$ on I_i and $\varphi_2 = \max \varphi\restriction_{I_i}$ on I_i. By definition of Riemann integrability, the partition can be chosen such that

$$\int_{S^1} \varphi(\phi)\, d\phi - \epsilon \le \int_{S^1} \varphi_1(\phi)\, d\phi \le \int_{S^1} \varphi_2(\phi)\, d\phi. \le \int_{S^1} \varphi(\phi)\, d\phi + \epsilon.$$

This implies that

$$\begin{aligned}(4.1.8)\quad &\int_{S^1} \varphi(\phi)\, d\phi - \epsilon \le \int_{S^1} \varphi_1(\phi)\, d\phi = \lim_{n\to\infty} \mathcal{B}_n(\varphi_1) \le \liminf_{n\to\infty} \mathcal{B}_n(\varphi) \\ &\le \limsup_{n\to\infty} \mathcal{B}_n(\varphi) \le \lim_{n\to\infty} \mathcal{B}_n(\varphi_2) = \int_{S^1} \varphi_2(\phi)\, d\phi \le \int_{S^1} \varphi(\phi)\, d\phi + \epsilon.\end{aligned}$$

Letting $\epsilon \to 0$ gives the result. □

Remark 4.1.16 The condition of Riemann integrability is essential. To see this, take a point x_0 and define the set A as the union of the arcs of length 2^{-k+2} centered at $R_\alpha^k(x_0)$ for $k \ge 0$. Although some of these arcs overlap, A is a union of arcs the sum of whose lengths is less than $1/2$, whereas $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \chi_A(R_\alpha^k(x)) = 1$. Of course, χ_A is not Riemann integrable.

2. Time Average and Space Average. The quantities on either side of (4.1.7) are averages.

Definition 4.1.17 Given a function φ, we call

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi\big(R_\alpha^k(x)\big)$$

its *time average* as sampled by following the orbit of x under the iterates of the rotation R_α. (Figure 4.1.5 illustrates this for $\varphi = \chi_{(0,1/2)}$.) The integral $\int_{S^1} \varphi(\phi)\, d\phi$ is called the *space average* of the function φ.

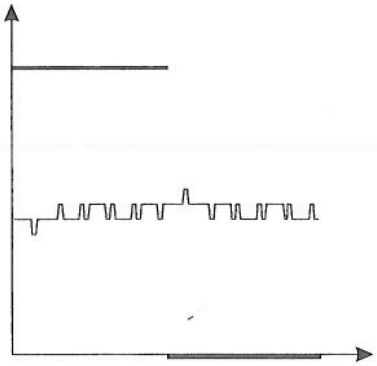


Figure 4.1.5. Time average.

These notions are both borrowed from physics, which is concerned with the measurement of observable quantities associated with a dynamical system. This means that there is a (measurable) quantity associated with the dynamical system in question that varies with the state of the dynamical system – in other words, one has a function defined on phase space whose value at a particular state of the dynamical system is displayed by the measuring device. Especially for systems that behave in unpredictable ways, it is quite natural to take a large number of successive measurements and average them. The limit of these averages is exactly the time average for the initial condition at which the measurements were begun.

The space average is more likely obtained as a result of calculations with a mathematical model of the physical system at hand. If one knows, as we do in our simple example, that space averages and time averages are supposed to coincide for the model one is testing, then the space average constitutes a prediction of the time average that is being measured, thus providing a means of verifying or falsifying the proposed model.

Returning to our situation, we note that the preceding result says that for any Riemann-integrable function the time average exists for the orbit of any point x and always coincides with the space average. This important property of irrational rotations is equivalent to uniform distribution and is referred to as *unique ergodicity*. This notion can be defined in the abstract setting of a continuous map of a compact metric space, even though there is no notion of an integral.

Definition 4.1.18 If X is a compact metric space and $f\colon X \to X$ a continuous map, then f is said to be *uniquely ergodic* if

$$\frac{1}{n}\sum_{k=0}^{n-1}\varphi(f^k(x))$$

converges to a constant uniformly (in x) for every continuous function φ.

4.1.6 The Kronecker–Weyl Method

In our arguments step functions played a special role. One can prove unique ergodicity of an irrational rotation in a much simpler, yet less elementary, way by using trigonometric polynomials to approximate continuous functions. This is possible due to the classical theorem of Weierstraß that says that continuous functions are uniform limits of trigonometric polynomials. This theorem is a close counterpart of a more familiar Weierstraß theorem that deals with the uniform approximation of a continuous function on an interval by polynomials. In this argument it is more convenient to use complex-valued functions.

Alternate proof of Proposition 4.1.14 Define the *characters* $c_m(x) := e^{2\pi imx} = \cos 2\pi mx + i \sin 2\pi mx$. If $m \neq 0$, then

$$c_m(R_\alpha(x)) = e^{2\pi im(x+\alpha)} = e^{2\pi im\alpha}e^{2\pi imx} = e^{2\pi im\alpha}c_m(x)$$

and

$$\left|\frac{1}{n}\sum_{k=0}^{n-1}c_m\big(R_\alpha^k(x)\big)\right| = \left|\frac{1}{n}\sum_{k=0}^{n-1}e^{2\pi imk\alpha}\right| = \frac{|1-e^{2\pi imn\alpha}|}{n|1-e^{2\pi im\alpha}|} \leq \frac{2}{n|1-e^{2\pi im\alpha}|} \to 0$$

as $n \to \infty$, because $\sum_{k=0}^{n} x^k = (1-x^{n+1})/(1-x)$.

Birkhoff averaging operators are linear; so if $p(x) = \sum_{i=-l}^{l} a_i c_i(x)$ is a trigonometric polynomial, then $\lim_{n\to\infty} \mathcal{B}_n(p)(x)$ exists and is constant. It is a_0 because this constant has to be the integral of p over S^1 (the operators $\mathcal{B}_n$ do not change the integral). The same arguments as above allow us to pass to uniform limits of trigonometric polynomials, that is, all continuous functions. □

This argument is more analytic and involves a much more straightforward calculation than the proof using step functions. Notice, however, that it does not give the original uniform distribution statement (Proposition 4.1.7), since characteristic functions are obviously not the uniform limit of trigonometric polynomials. To obtain uniform distribution for intervals from that for functions one can use the argument from the proof of Theorem 4.1.15 backwards: Approximate χ_A by continuous functions $\varphi_1 \leq \chi_A \leq \varphi_2$ such that $\int(\varphi_2 - \varphi_1) < \epsilon$ (Figure 4.1.6) and repeat the calculation (4.1.8).

4.1.7 Group Translations

Irrational rotations serve as the starting point for a number of fruitful generalizations. Let us discuss one of them. The circle is a compact abelian *group*, and the rotation can be represented in group terms as the group

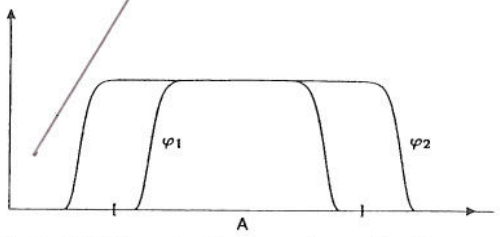


Figure 4.1.6. Approximation by continuous functions.

multiplication or translation

$$L_{g_0}: G \to G, \quad L_{g_0} g = g_0 g.$$

The orbit of the unit element $e \in G$ is the cyclic subgroup $\{g_0^n\}_{n\in\mathbb{Z}}$. Proposition 4.1.1 is closely related to the fact that the circle does not have proper infinite closed subgroups. To say that an orbit is dense requires a notion of approximation, so we define a *topological group* to be a group with a metric for which every L_g is a homeomorphism and taking inverses is continuous.

Proposition 4.1.19 *If the translation L_{g_0} on a topological group G is topologically transitive, then it is minimal.*

Proof For $g, g' \in G$ denote by $A, A' \subset G$ the closures of the orbits of g and g', respectively. Now $g_0^n g' = g_0^n g(g^{-1}g')$, so $A' = Ag^{-1}g'$ and $A' = G$ if and only if $A = G$. □

EXERCISES

Exercise 4.1.1 Prove that for the metric $d(x, y) := \min\{|b - a| \mid a \in x, b \in y\}$ on the set $X = \mathbb{R}/\mathbb{Z} := \{[x] \mid x \in \mathbb{R}\}$ every rotation is an isometry (as in Definition A.1.16).

Exercise 4.1.2 Take $c = 7.1$ in Section 4.1.3 and determine the next closest return.

Exercise 4.1.3 Prove the properties in Remark 4.1.11.

Exercise 4.1.4 For a rotation R_α find $N \in \mathbb{N}$ in terms of α such that $F_{(0.1/2)}(x, n)/n \geq 0.45$ for all $n \geq N$ (see Section 4.1.4).

Exercise 4.1.5 Suppose the motion of the sun and moon as observed from a specific place on earth are strictly periodic and that the time difference between sunrise and moonrise is never twice the same. Prove that this difference is uniformly distributed.

Exercise 4.1.6 Give an example of a homeomorphism of a complete metric space that has a dense orbit but no dense semiorbit.

Exercise 4.1.7 Give an example of a homeomorphism of a compact metric space that has a dense orbit but no dense semiorbit.

Exercise 4.1.8 Prove that two minimal sets (Definition 4.1.4) are either disjoint or equal.

Exercise 4.1.9 Prove that a contracting map of a compact space is uniquely ergodic

Exercise 4.1.10 Give an example of a continuous map f of a compact metric space X such that

$$\frac{1}{n}\sum_{k=0}^{n-1} \varphi(f^k(x))$$

converges uniformly (in x) for every continuous function φ, but f is not uniquely ergodic.

Exercise 4.1.11 Using enough digits of the decimal expansion of π, find the classical approximations 21/7 and 355/113 and write the result in the form described in Section 4.1.3. Find the fourth term in the continued-fraction approximation and explain how the size of this number is reflected in the quality of the approximation.

PROBLEMS FOR FURTHER STUDY

Problem 4.1.12 Let G be a metrizable compact topological group. Suppose for some $g_0 \in G$ the translation L_{g_0} is topologically transitive. Prove that G is abelian.

Problem 4.1.13 Show that a finite abelian group has a uniquely ergodic translation if and only if it is cyclic.

Problem 4.1.14 Prove that the circle map $x \mapsto x + (1/4)\sin^2 \pi x$ shown in Figure 2.2.4 is uniquely ergodic.

Problem 4.1.15 Define the following metric d_2 on the group $\mathbb{Z}$ of all integers: $d_2(m, n) = \|m - n\|_2$, where

$$\|n\|_2 = 2^{-k} \qquad \text{if } n = 2^k l \text{ with an odd number } l.$$

The completion of $\mathbb{Z}$ with respect to that metric is called the group of *2-adic* or *dyadic integers* and is usually denoted by $\mathbb{Z}_2$. It is a compact topological group. Let $\mathbb{Z}_2^+$ be the closure of the even integers with respect to the metric d_2. $\mathbb{Z}_2^+$ is a subgroup of $\mathbb{Z}_2$ of index two.

Prove that for $g_0 \in \mathbb{Z}_2$ the translation $L_{g_0}: \mathbb{Z}_2 \to \mathbb{Z}_2$ is topologically transitive if and only if $g_0 \in \mathbb{Z}_2 \setminus \mathbb{Z}_2^+$.

This is an example of a class of systems called *adding machines*. An equivalent description is given in Definition 11.3.10, and Theorem 11.3.11 shows that this dynamical system is a subsystem of the quadratic map $f_\lambda: [0, 1] \to [0, 1]$, $f_\lambda(x) := \lambda x(1 - x)$ from Section 2.5 for a particular value of λ.

4.2 SOME APPLICATIONS OF DENSITY AND UNIFORM DISTRIBUTION

There are numerous situations in which one would like to obtain information of some asymptotic nature and where the dynamics of circle rotations or toral translations, which are considered in the next chapter, plays a role, possibly behind the scenes, that makes it possible to obtain this asymptotic information from the knowledge we have acquired so far. In this section we show some such examples.

4.2.1 Distribution of Values for Periodic Functions

Let $(x_n)_{n\in\mathbb{N}}$ be a sequence of real numbers. A natural way to describe the distribution of values of such a sequence would be to consider the asymptotic frequencies with which this sequence "visits" various intervals.

Definition 4.2.1 Given a sequence $(x_n)_{n\in\mathbb{N}}$ and $a < b$, let $F_{a,b}(n)$ be the number of integers k such that $1 \leq k \leq n$ and $a < x_k < b$. We say that $(x_n)_{n\in\mathbb{N}}$ has an *asymptotic distribution* if for any a, b, $-\infty \leq a < b \leq \infty$ the limit

$$\lim_{n\to\infty} \frac{F_{a,b}(n)}{n}$$

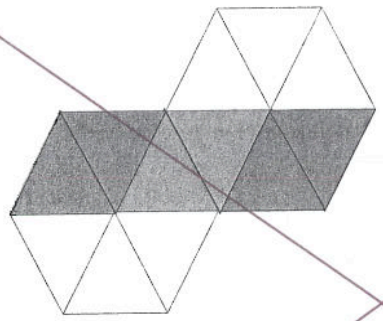


Figure 4.2.15. Parallelogram.

Energy conservation reduces this to three dimensions, and conservation of the size of angular momentum to two. These two dimensions are parametrized by a time parameter along an ellipse and a perihelion angle. This is therefore a system similar to the mathematical pendulum (Section 6.2.2) where one gets flows on circles, that is, one-dimensional invariant tori. However, the Kepler problem for several planets *without mutual interaction* gives higher-dimensional invariant tori with linear flows on them. This is the central feature of *complete integrability* in Hamiltonian dynamical systems.

■ EXERCISES

■ **Exercise 4.2.1** Give a detailed proof of (4.2.1).

■ **Exercise 4.2.2** Verify directly that for any fixed number m the sum of $\lg(p+1) - \lg p$ over all p with exactly m digits is 1, as it should be according to Proposition 4.2.7.

■ **Exercise 4.2.3** Verify the calculation needed to deduce Proposition 4.2.7 from Proposition 4.1.7 or Theorem 4.2.3.

■ **Exercise 4.2.4** Referring to Proposition 4.2.7, determine $\lim_{n\to\infty} F_{10}^2(n)/n$ and find the asymptotic frequencies of 0 and 9, respectively, as the *second* digit of powers of 2.

■ **Exercise 4.2.5** Referring to the proof of Proposition 4.2.8, assume $\gamma \neq 0$ and replace the section C_1 by the section $C_2 := \{x_2 = 0\}$. Prove that the resulting return map is a rotation and determine the rotation angle in terms of γ.

■ **Exercise 4.2.6** Verify by direct calculation of the time derivatives that the functions $x_1^2 + x_2^2$ and $x_3^2 + x_4^2$ are invariant under (4.2.5).

■ **Exercise 4.2.7** Formulate the natural uniform distribution property referred to in Proposition 4.2.9 and proved in Section 4.2.5.4.

■ **Exercise 4.2.8** Prove that any closed proper subgroup Γ of $\mathbb{R}$ is cyclic, that is, $\Gamma = \{na\}_{n\in\mathbb{Z}}$ for some $a \in \mathbb{R}$.

■ **Exercise 4.2.9** Given an initial direction, how many slopes are there for the billiard flow in a square and in each of the two triangles, and what are they?

■ **Exercise 4.2.10** Suppose a horizontal light beam enters a circular room with mirrored walls. Describe the possibilities for which areas of the room will be best lit.

■ **Exercise 4.2.11** Prove that a complete unfolding of a regular pentagon covers every point of the plane infinitely many times.

■ **Exercise 4.2.12** Obtain the continuation of orbits in the billiard description of the 2-particle system by interpreting double collisions as limits of a series of simple collisions.

4.3 INVERTIBLE CIRCLE MAPS

The success in analyzing circle rotations is due in large part to the fact that these come from linear dynamical systems, namely, from rotations of the plane (Section 3.1). This causes the great homogeneity of the orbit structure that gives uniform density of orbits and uniform distribution. However, another ingredient, perhaps less apparent, is the simple structure of the circle itself. Analogously to the study of interval homeomorphisms (Section 2.3.1) this makes it possible to give a fairly satisfactory analysis of the orbit structure of any invertible map of the circle. One-dimensionality of the circle provides two (related) features that make a fairly detailed analysis possible: the (cyclic) ordering of its points and the Intermediate-Value Theorem. These have the effect of tying together different orbits tightly enough to make the possible orbit structures relatively easy to describe. The importance of the order structure will become particularly apparent in Proposition 4.3.11 and Proposition 4.3.15.

For noninvertible maps of an interval or of the circle the order of points may not be preserved and hence use of this first property fails, while the Intermediate-Value Theorem can still be used so long as we have continuity. Accordingly, the structural features are much more complicated while still amenable to rather extensive analysis. Chapter 11 outlines this for some interval maps.

One principle that will manifest itself in various guises throughout this section is that while, unlike the situation with rotations, the orbit structure of invertible circle maps is not always entirely homogeneous, the asymptotic behavior is in various different ways about as homogeneous, or at least coherent, as the entire orbit structure of a rotation and, in fact, ultimately turns out to look much like a rotation.

In this section a fundamental dichotomy is central: A circle homeomorphism (Definition A.1.16) may or may not have periodic points. Every orbit has the same type of asymptotic behavior, and it corresponds in a precise sense to the behavior of an orbit of a rational or an irrational rotation, respectively. The tool that leads to this conclusion is a parameter that reflects asymptotic rotation rates and is rational or not according to whether there are periodic points.

4.3.1 Lift and Degree

Recall the relation between the circle $S^1 = \mathbb{R}/\mathbb{Z}$ and the line $\mathbb{R}$ (see Section 2.6.2). There is a projection $\pi : \mathbb{R} \to S^1$, $x \mapsto [x]$, where $[x]$ is the equivalence class of x in

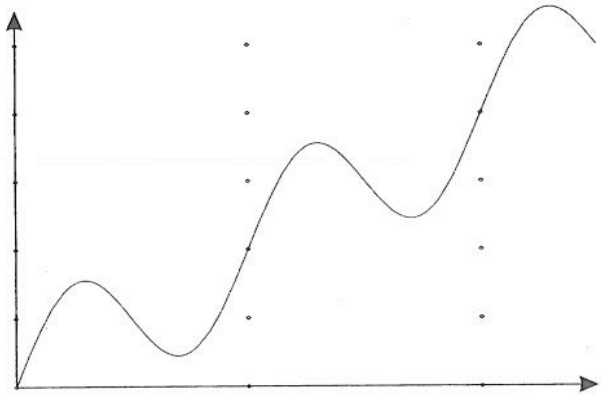


Figure 4.3.1. A lift and degree.

$\mathbb{R}/\mathbb{Z}$ as in Section 2.6.2. Here $[\cdot]$ denotes an equivalence class, whereas the integer part of a number is written $\lfloor\cdot\rfloor$. We use $\{\cdot\}$ for the fractional part.

Proposition 4.3.1 *If $f\colon S^1 \to S^1$ is continuous, then there exists a continuous $F\colon \mathbb{R} \to \mathbb{R}$, called a* lift *of f to $\mathbb{R}$, such that*

$$f \circ \pi = \pi \circ F, \tag{4.3.1}$$

that is, $f([z]) = [F(z)]$. Such a lift is unique up to an additive integer constant, and $\deg(f) := F(x+1) - F(x)$ is an integer independent of $x \in \mathbb{R}$ and the lift F. It is called the degree *of f. If f is a homeomorphism, then $|\deg(f)| = 1$.*

Proof *Existence:* Pick a point $p \in S^1$. Then $p = [x_0]$ for some $x_0 \in \mathbb{R}$ and $f(p) = [y_0]$ for some $y_0 \in \mathbb{R}$. From these choices of x_0 and y_0 define $F\colon \mathbb{R} \to \mathbb{R}$ by requiring that $F(x_0) = y_0$, F is continuous, and $f([z]) = [F(z)]$ for all $z \in \mathbb{R}$. One can construct such an F by varying the initial point p continuously, which causes $f(p)$ to vary continuously. Then there is no ambiguity of how to vary x and y continuously, and thus $F(x) = y$ defines a continuous map.[3]

Uniqueness: Suppose $\tilde{F}$ is another lift. Then $[\tilde{F}(x)] = f([x]) = [F(x)]$ for all x, meaning $\tilde{F} - F$ is always an integer. Because it is continuous it must be constant.

Degree: $F(x+1) - F(x)$ is an integer (now evidently independent of the choice of lift) because $[F(x+1)] = f([x+1]) = f([x]) = [F(x)]$. By continuity, $F(x+1) - F(x) =: \deg(f)$ must be a constant.

Invertibility: If $\deg(f) = 0$, then $F(x+1) = F(x)$ and thus F is not monotone. Then f is noninvertible because it cannot be monotone. If $|\deg(f)| > 1$, then $|F(x+1) - F(x)| > 1$ and, by the Intermediate-Value Theorem, there exists a

[3] To elaborate, take $\delta > 0$ such that $d([x],[x']) \le \delta$ implies $d(f([x]), f([x'])) < 1/2$. Then define F on $[x_0 - \delta, x_0 + \delta]$ as follows: If $|x - x_0| \le \delta$, then $d(f([x]), q) < 1/2$ and there is a unique $y \in (y_0 - 1/2, y_0 + 1/2)$ such that $[y] = f([x])$. Define $F(x) = y$. Analogous steps extend the domain by another δ at a time, until F is defined on an interval of unit length. Then $f([z]) = [F(z)]$ defines F on $\mathbb{R}$.

$y \in (x, x+1)$ with $|F(y) - F(x)| = 1$. Then $f([y]) = f([x])$ and $[y] \neq [x]$, so f is noninvertible. □

Definition 4.3.2 Suppose f is invertible. If $\deg(f) = 1$, then we say that f is orientation-preserving; if $\deg(f) = -1$, then f is said to reverse orientation.

Remark 4.3.3 The function $F(x) - x\deg(f)$ is periodic because

$$F(x+1) - (x+1)\deg(f) = F(x) + \deg(f) - (x+1)\deg(f) = F(x) + x\deg(f)$$

for all x. In particular, if f is an orientation-preserving homeomorphism, then $F(x) - x$ is periodic and so $F - \mathrm{Id}$ is bounded. A slightly stronger observation will come in handy soon.

Lemma 4.3.4 *If f is an orientation-preserving circle homeomorphism and F a lift, then $F(y) - y \le F(x) - x + 1$ for all $x, y \in \mathbb{R}$.*

Proof Let $k = \lfloor y - x \rfloor$. Then

$$\begin{aligned} F(y) - y &= F(y) + F(x+k) - F(x+k) + (x+k) - (x+k) - y \\ &= (F(x+k) - (x+k)) + (F(y) - F(x+k)) - (y - (x+k)). \end{aligned} \tag{4.3.2}$$

Now $F(x+k) - (x+k) = F(x) - x$ and $0 \le y - (x+k) < 1$ by choice of k, so $F(y) - F(x+k) \le 1$. Thus the right-hand side above is at most $F(x) - x + 1 - 0$. □

4.3.2 Rotation Number

The presence or absence of periodic points is determined by a single parameter called the rotation number. It also tells us which rotation to compare a circle homeomorphism to.

Proposition 4.3.5 *Let $f\colon S^1 \to S^1$ be an orientation-preserving homeomorphism and $F\colon \mathbb{R} \to \mathbb{R}$ a lift of f. Then*

$$\rho(F) := \lim_{|n|\to\infty} \frac{1}{n}(F^n(x) - x) \tag{4.3.3}$$

exists for all $x \in \mathbb{R}$. $\rho(F)$ is independent of x and well defined up to an integer; that is, if $\tilde{F}$ is another lift of f, then $\rho(F) - \rho(\tilde{F}) = F - \tilde{F} \in \mathbb{Z}$. $\rho(F)$ is rational if and only if f has a periodic point.

The fact that the rotation number is independent of the point is the first manifestation of the coherent asymptotic behavior of orbits that we will come to expect. This proposition justifies the following terminology:

Definition 4.3.6 $\rho(f) := [\rho(F)]$ is called the *rotation number* of f.

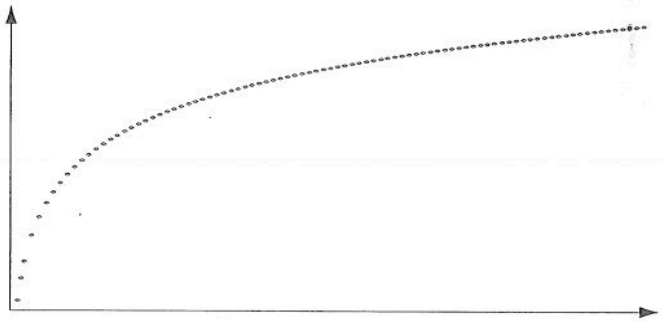


Figure 4.3.2. Subadditivity.

A sequence $(a_n)_{n\in\mathbb{N}}$ with $a_{n+m} \le a_n + a_m$ is said to be *subadditive*. Existence of the rotation number is due to a similar property of the right-hand side of (4.3.3).

Lemma 4.3.7 *If a sequence $(a_n)_{n\in\mathbb{N}}$ satisfies $a_{m+n} \le a_n + a_{m+k} + L$ for all $m, n \in \mathbb{N}$ and some k and L, then $\lim_{n\to\infty} a_n/n \in \mathbb{R} \cup \{-\infty\}$ exists.*

Proof $a_{m+k} \le a_m + a_{2k} + L$ gives $a_{m+n} \le a_m + a_n + a_{2k} + 2L = a_m + a_n + L'$, so we may take $k = 0$. Let $a := \underline{\lim}_{n\to\infty} a_n/n \in \mathbb{R} \cup \{-\infty\}$.

If $a < b < c$ and $n > 2L/(c-b)$ such that $a_n/n < b$, then for any $l \ge n$ that satisfies $l(c-b) > 2\max_{r<n} a_r$ we can write $l = nk + r$ with $r < n$. This implies $a_l/l \le (ka_n + a_r + kL)/l \le a_n/n + a_r/l + (L/n) < c$, so $\overline{\lim}_{l\to\infty} a_l/l \le c$. Since $c > a$ was arbitrary, this proves the lemma. □

Proof of Proposition 4.3.5 *Independence of x*: Remark 4.3.3 gives $F(x+1) = F(x) + 1$. If $|x - y| < 1$, then $|F(y) - F(x)| < 1$ and

$$\left|\frac{1}{n}|F^n(x) - x| - \frac{1}{n}|F^n(y) - y|\right| \le \frac{1}{n}(|F^n(x) - F^n(y)| + |x - y|) \le \frac{2}{n}.$$

Thus the rotation numbers of x and y coincide, if one of them exists.

Existence: Take $x \in \mathbb{R}$ and $a_n := F^n(x) - x$. Then

$$a_{m+n} = F^{m+n}(x) - x = F^m(F^n(x)) - F^n(x) + a_n \le a_m + 1 + a_n$$

by Lemma 4.3.4 applied to f^m and F^m. Thus Lemma 4.3.7 shows that a_n/n converges, but possibly, to $-\infty$. However,

$$\frac{a_n}{n} = \frac{1}{n}\sum_{i=0}^{n-1}(F^{i+1}(x) - F^i(x)) = \frac{1}{n}\sum_{i=0}^{n-1}(F(x_i) - x_i) \ge \min F(y) - y,$$

so the limit is a real number $\rho(F)$.

Also, $\rho(F+m) = \lim_{|n|\to\infty}(1/n)(F^n(x) + nm - x) = \rho(F) + m$ for $m \in \mathbb{Z}$, that is, $\rho(F)$ is well defined (mod 1).

Periodic points: If f has a q-periodic point, then $F^q(x) = x + p$ for a lift x of it and some $p \in \mathbb{Z}$. If $m \in \mathbb{N}$, then

$$\frac{F^{mq}(x) - x}{mq} = \frac{1}{mq}\sum_{i=0}^{m-1} F^q(F^{iq}(x)) - F^{iq}(x) = \frac{mp}{mq} = \frac{p}{q};$$

so $\rho(F) = p/q$.

Conversely, for any lift F the definition of rotation number yields

$$\rho(F^m) = \lim_{n\to\infty}\frac{1}{n}((F^m)^n(x) - x) = m\lim_{n\to\infty}\frac{1}{mn}(F^{mn}(x) - x) = m\rho(F);$$

so if $\rho(f) = p/q \in \mathbb{Q}$, then $\rho(f^q) = 0$ since the rotation number is defined up to an integer. Therefore we need only show:

Claim If $\rho(f) = 0$, then f has a fixed point.

Suppose f has no fixed point and let F be a lift such that $F(0) \in [0, 1)$. Then $F(x) - x \notin \mathbb{Z}$ for all $x \in \mathbb{R}$ since $F(x) - x \in \mathbb{Z}$ would imply that $[x]$ is a fixed point for f. Therefore, $0 < F(x) - x < 1$ for all $x \in \mathbb{R}$. Since $F - \mathrm{Id}$ is continuous and periodic, it attains its minimum and maximum and therefore there exists a $\delta > 0$ such that

$$0 < \delta \le F(x) - x \le 1 - \delta < 1$$

for all $x \in \mathbb{R}$. In particular, we can take $x = F^i(0)$ and use $F^n(0) = F^n(0) - 0 = \sum_{i=0}^{n-1} F^{i+1}(0) - F^i(0)$ to get

$$n\delta \le F^n(0) \le (1-\delta)n$$

or

$$\delta \le \frac{F^n(0)}{n} \le 1 - \delta.$$

As $n \to \infty$, this gives $\rho(F) \ne 0$, which proves the claim by contraposition. □

All periodic orbits, if any, have the same period:

Proposition 4.3.8 *Let $f\colon S^1 \to S^1$ be an orientation-preserving homeomorphism. Then all periodic orbits have the same period.*

In fact, if $\rho(f) = [p/q]$ with $p, q \in \mathbb{Z}$ relatively prime, then the lift F of f, with $\rho(F) = p/q$ satisfies $F^q(x) = x + p$ whenever $[x]$ is a periodic point, that is, the set of periodic points of f lifts to the set of fixed points of $F^q - \mathrm{Id} - p$.

Proof If $[x]$ is a periodic point, then $F^r(x) = x + s$ for some $r, s \in \mathbb{Z}$ and

$$\frac{p}{q} = \rho(F) = \lim_{n\to\infty}\frac{F^n(x) - x}{nr} = \lim_{n\to\infty}\frac{ns}{nr} = \frac{s}{r}.$$

This means that $s = mp$ and $r = mq$ and that therefore $F^{mq}(x) = x + mp$.

Claim $F^q(x) = x + p$.

If $F^q(x) - p > x$, then monotonicity of F implies

$$F^{2q}(x) - 2p = F^q(F^q(x) - p) - p \ge F^q(x) - p > x$$

and inductively $F^{mq}(x) - mp > x$, which is impossible. Likewise, $F^q(x) - p < x$ is impossible because it implies $F^{mq}(x) - mp < x$. This proves the claim. □

4.3.3 Conjugacy Invariance

The notion of topological conjugacy is central to many aspects of dynamics and will be introduced later (Definition 7.3.3). The rotation number provides the first nontrivial example of a conjugacy invariant, due to the following result:

Proposition 4.3.9 *If $f, h\colon S^1 \to S^1$ are orientation-preserving homeomorphisms, then $\rho(h^{-1}fh) = \rho(f)$.*

Proof Let F and H be lifts of f and h, respectively, that is, $\pi F = f\pi$ and $\pi H = h\pi$. Then $\pi H^{-1} = h^{-1}h\pi H^{-1} = h^{-1}\pi H H^{-1} = h^{-1}\pi$, so H^{-1} is a lift of h^{-1}. Also, $H^{-1}FH$ is a lift of $h^{-1}fh$ since $\pi H^{-1}FH = h^{-1}\pi FH = h^{-1}f\pi H = h^{-1}fh\pi$.

Suppose H is such that $H(0) \in [0, 1)$. We need to estimate

$$|H^{-1}F^nH(x) - F^n(x)| = |(H^{-1}FH)^n(x) - F^n(x)|.$$

(1) For $x \in [0, 1)$ we have $0 - 1 < H(x) - x < H(x) < H(1) < 2$, and by periodicity $|H(x) - x| < 2$ for $x \in \mathbb{R}$. Similarly, $|H^{-1}(x) - x| < 2$ for $x \in \mathbb{R}$.

(2) If $|y - x| < 2$, then $|F^n(y) - F^n(x)| < 3$ since $|[y] - [x]| \le 2$ and thus

$$\begin{aligned} -3 \le [y] - [x] - 1 &= F^n([y]) - F^n([x] + 1) < F^n(y) - F^n(x) \\ &< F^n([y] + 1) - F^n([x]) = [y] + 1 - [x] \le 3. \end{aligned}$$

Those two estimates yield

$$|H^{-1}F^nH(x) - F^n(x)| \le |H^{-1}F^nH(x) - F^nH(x)| + |F^nH(x) - F^n(x)| < 2 + 3,$$

so $|(H^{-1}FH)^n(x) - F^n(x)|/n < 5/n$ and $\rho(H^{-1}FH) = \rho(F)$ by (4.3.3). □

The behavior of the rotation number under orientation-reversing conjugacies is the subject of Exercise 4.3.6.

4.3.4 Circle Homeomorphisms with Periodic Points

The orbit structure of a circle homeomorphism can be described in a fairly complete fashion. We first do this for the case with periodic points.

The first level of description is that each periodic orbit is ordered in the same way as those of the corresponding rotation. This means that the periodic orbits of an orientation-preserving circle homeomorphism behave like those of the circle rotation with the same rotation number. So not only is there an internal coherence of the various periodic orbits as described by Proposition 4.3.8, but they also are qualitatively compatible with those of a rotation. This was, in fact, presaged by the proof of Proposition 4.3.8.

Before proving this, the "ordering" of an orbit has to be defined. It is the sequence in which one encounters the points of the orbit when moving from its initial point in the positive direction. Formally, one can define this using lifts:

Definition 4.3.10 Given $x_0, \dots, x_{n-1} \in S^1$, take $\tilde{x}_0, \dots, \tilde{x}_{n-1} \in [\tilde{x}_0, \tilde{x}_0 + 1) \subset \mathbb{R}$ such that $[\tilde{x}_i] = x_i$. Then the *ordering* of $(x_0, \dots, x_{n-1})$ on S^1 is the permutation σ of $\{1, \dots, n-1\}$ such that $\tilde{x}_0 < \tilde{x}_{\sigma(1)} < \cdots < \tilde{x}_{\sigma(n-1)} < \tilde{x}_0 + 1$.

As a warmup, we find the ordering σ of $\pi(\{0, p/q, 2p/q, \dots, (q-1)p/q\})$ on S^1, to which we later compare that of a periodic orbit. Define $k \in \mathbb{N}$ by $0 < k < q$ and $kp \equiv 1 \pmod q$. Then k minimizes the fractional part $\{ip/q\}$ for $0 < i < q$ and hence $k = \sigma(1)$. Inductively, $ki \equiv \sigma(i) \pmod q$. This defines the ordering σ of

$$\pi\left(\left\{0, \frac{p}{q}, \frac{2p}{q}, \dots, \frac{(q-1)p}{q}\right\}\right).$$

Therefore, we want to prove:

Proposition 4.3.11 *Let $f\colon S^1 \to S^1$ be an orientation-preserving homeomorphism. Suppose p and q are relatively prime and there is an $x \in S^1$ such that $f^q(x) = x$. Then the ordering of $\{x, f(x), f^2(x), \dots, f^{q-1}(x)\}$ on S^1 is given by $\sigma(i) = ki \pmod q$, where $kp \equiv 1 \pmod q$.*

Proof Fix $\tilde{x} \in \pi^{-1}([x])$ and a lift F of f such that $F^q(\tilde{x}) = \tilde{x} + p$ (Proposition 4.3.8). Then $[\tilde{x}, \tilde{x} + p]$ is partitioned (up to common endpoints) into $p \cdot q$ subintervals by $A := \pi^{-1}(\{x, f(x), f^2(x), \dots, f^{q-1}(x)\})$, and into q subintervals $I_i = [F^i(\tilde{x}), F^{i+1}(\tilde{x})]$, $i = 0 \dots q-1$. Since F is a bijection between any I_i and I_{i+1} and preserves A, each I_i contains $p+1$ points of A. Take $k, r \in \mathbb{Z}$ such that the right neighbor of $\tilde{x}$ in A is $\tilde{x}_1 = F^k(\tilde{x}) - r$. Since $\bar{F} = F^k - r$ is increasing on $\mathbb{R}$ and preserves A, the facts that $\tilde{x}_1 = \bar{F}(\tilde{x})$ is the nearest right neighbor of $\tilde{x}$ in A and that $[\tilde{x}, F(\tilde{x})]$ is divided into p subintervals by A show that $\bar{F}^p(\tilde{x}) = F(\tilde{x})$ and hence $f^{kp}(x) = f(x)$. Therefore k is the unique integer between 0 and $q-1$ such that $kp \equiv 1 \pmod q$, and the ordering of the orbit $\{x, f(x), f^2(x), \dots, f^{q-1}(x)\}$ is given by $ki \equiv \sigma(i) \pmod q$. □

The next proposition says that for circle homeomorphisms with rational rotation number all nonperiodic orbits are asymptotic to periodic orbits. This yields a complete classification of possible orbits with rational rotation numbers.

Proposition 4.3.12 *Let $f\colon S^1 \to S^1$ be an orientation-preserving homeomorphism with rational rotation number $\rho(f) = p/q \in \mathbb{Q}$. Then there are two possible types of nonperiodic orbits for f:*

(1) *If f has exactly one periodic orbit, then every other point is heteroclinic under f^q to two points on the periodic orbit (Definition 2.3.4). These points are different if the period is greater than one. (If the period is one, then all orbits are* homoclinic *to the fixed point, as shown in Figure 4.3.3.)*

(2) *If f has more than one periodic orbit, then each nonperiodic point is heteroclinic under f^q to two points on different periodic orbits.*

Proof We can identify f^q with a homeomorphism of an interval by taking a lift z of a fixed point of f^q and restricting a lift $F^q(\cdot) - p$ of f to $[z, z+1]$. Then the statement follows from Proposition 2.3.5 applied to this interval map, except for the last part of (2), that the two periodic orbits in question

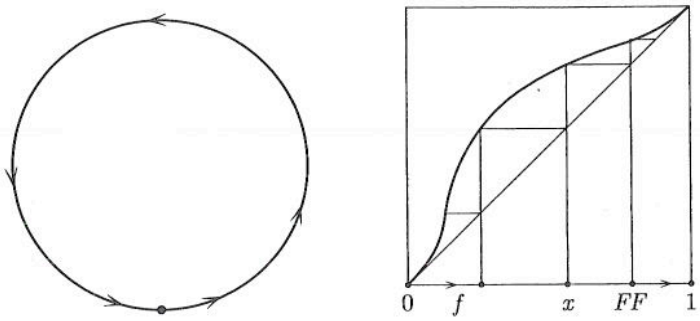


Figure 4.3.3. A semistable point.

are different. But if there is an interval $I = [a, b] \subset \mathbb{R}$ such that a and b are adjacent zeros of $F^q - \mathrm{Id} - p$ and a, b project to the same periodic orbit, then f has only one periodic orbit because, if $[a] = x \in S^1$, $[b] = f^k(x) \in S^1$, then $\bigcup_{n=0}^{q-1} f^{nk}(\pi((a, b)))$ covers the complement of $\{f^n(x)\}_{n=0}^{q-1}$ in S^1 and contains no periodic points. By invariance, $f^{nk}(\pi((a, b)))$ does not either. □

Remark 4.3.13 This means that the asymptotic behavior is highly coherent for all orbits, not only periodic ones, and also coherent with the structure of the corresponding rotation.

As a particular case, if there is only one periodic orbit, then it is semistable. It "repels on one side and attracts on the other", as, for example, the point $x = 0$ under the diffeomorphism $f\colon S^1 \to S^1$ induced by the map

$$x \mapsto x + \frac{1}{4}\sin^2 \pi x \pmod 1.$$

Nonperiodic points are not just individually asymptotic to periodic points, but this behavior is coherent for iterates of points under f; so for a nonperiodic point x the points $x, f(x), \ldots, f^{q-1}(x)$ are all forward asymptotic to the corresponding iterate $y, f(y), \ldots, f^{q-1}(y)$ of a periodic point, and they are moving in the same direction. This follows immediately from monotonicity (compare Lemma 2.3.2):

Lemma 4.3.14 *If $I \subset \mathbb{R}$ is an interval whose endpoints are adjacent zeros of $F^q - \mathrm{Id} - p$, then $F^q - \mathrm{Id} - p$ has the same sign on the interiors of I and $F(I)$.*

Proof If $F^q - \mathrm{Id} - p > 0$ on I, then $F^q(x) > x + p$ for all $x \in I$, and monotonicity of F implies $F^q(F(x)) = F(F^q(x)) > F(x + p) = F(x) + p$ for all $x \in I$. Therefore $F^q - \mathrm{Id} - p > 0$ on $F(I)$.

The case of $F^q - \mathrm{Id} - p < 0$. is similar. □

Thus for a circle homeomorphism with a periodic point all orbits are asymptotically periodic with the same period and in a coherent way.

4.3.5 Circle Homeomorphisms Without Periodic Points

We show, analogously to Proposition 4.3.11, that the orbits of a circle homeomorphism without periodic points are ordered as those for the corresponding rotation.

Proposition 4.3.15 *Let $F\colon \mathbb{R} \to \mathbb{R}$ be a lift of an orientation-preserving homeomorphism $f\colon S^1 \to S^1$ with $\rho := \rho(F) \notin \mathbb{Q}$. Then, for $n_1, n_2, m_1, m_2 \in \mathbb{Z}$ and $x \in \mathbb{R}$,*

$$n_1\rho + m_1 < n_2\rho + m_2 \quad \textit{if and only if} \quad F^{n_1}(x) + m_1 < F^{n_2}(x) + m_2.$$

The left of these inequalities is the special case of the one on the right when F is the rotation by ρ.

Proof We do not have equality on the right for any x because this would imply $F^{n_1}(x) - F^{n_2}(x) \in \mathbb{Z}$, and hence that $[x]$ is periodic. Thus, for given $n_1, n_2, m_1, m_2 \in \mathbb{Z}$, the continuous expression $F^{n_1}(x) + m_1 - F^{n_2}(x) - m_2$ never changes sign and the second inequality is independent of x.

Now assume $F^{n_1}(x) + m_1 < F^{n_2}(x) + m_2$ for all x. Substituting $y := F^{n_2}(x)$ shows that this is equivalent to

$$F^{n_1-n_2}(y) - y < m_2 - m_1 \qquad \text{for all } y \in \mathbb{R}.$$

In particular, for $y = 0$ we get $F^{n_1-n_2}(0) < m_2 - m_1$, and $y = F^{n_1-n_2}(0)$ gives

$$F^{2(n_1-n_2)}(0) < (m_2 - m_1) + F^{n_1-n_2}(0) < 2(m_2 - m_1).$$

Inductively, $F^{n(n_1-n_2)}(0) < n(m_2 - m_1)$ and

$$\rho = \lim_{n\to\infty} \frac{F^{n(n_1-n_2)}(0)}{n(n_1 - n_2)} < \lim_{n\to\infty} \frac{n(m_2 - m_1)}{n(n_1 - n_2)} = \frac{m_2 - m_1}{n_1 - n_2}$$

(with strict inequality since $\rho \notin \mathbb{Q}$). Consequently, $n_1\rho + m_1 < n_2\rho + m_2$. This proves "if". Reversing all inequalities proves the converse. □

The preceding proposition bears some resemblance to the earlier result that *periodic* orbits are ordered like those for the corresponding rotation. It is stronger because it applies to every orbit, rather than a naturally distinguished subset.

This helps us in our study of the asymptotic behavior of orbits for homeomorphisms without periodic points.

Lemma 4.3.16 *Let $f\colon S^1 \to S^1$ be an orientation-preserving homeomorphism without periodic points, $m, n \in \mathbb{Z}$, $m \neq n$, $x \in S^1$, and $I \subset S^1$ a closed interval with endpoints $f^m(x)$ and $f^n(x)$. Then every semiorbit meets I.*

Remark 4.3.17 For $x \neq y \in S^1$ there are exactly two intervals in S^1 with endpoints x and y. The lemma holds for either choice. Since x is not periodic, I is not a point.

Proof Consider positive semiorbits $(f^n(y))_{n\in\mathbb{N}}$. The proof for negative semiorbits is exactly the same. To prove the lemma it suffices to show that the backward iterates of I cover S^1, that is, $S^1 \subset \bigcup_{k\in\mathbb{N}} f^{-k}(I)$.

Let $I_k := f^{-k(n-m)}(I)$ and note that these are all contiguous: If $k \in \mathbb{N}$, then I_k and I_{k-1} have a common endpoint. Consequently, if $S^1 \neq \bigcup_{k\in\mathbb{N}} I_k$, then the sequence of endpoints converges to some $z \in S^1$. But then

$$\begin{aligned} z &= \lim_{k\to\infty} f^{-k(n-m)}(f^m(x)) = \lim_{k\to\infty} f^{(-k+1)(n-m)}(f^m(x)) \\ &= \lim_{k\to\infty} f^{(n-m)}\big(f^{-k(n-m)}(f^m(x))\big) = f^{(n-m)}\big(\lim_{k\to\infty} f^{-k(n-m)}(f^m(x))\big) \\ &= f^{(n-m)}(z) \end{aligned}$$

is periodic, contrary to the assumption. □

If there are periodic points, they provide all the accumulation points of orbits. Now we see what set plays this role when the rotation number is irrational.

Definition 4.3.18 The set $\omega(x) := \bigcap_{n\in\mathbb{N}} \overline{\{f^i(x) \mid i \geq n\}}$ of accumulation points of the positive semiorbit of x is called the ω-limit set of x.

If there are periodic points, all ω-limit sets are periodic orbits. If there are no periodic points, the ω-limit sets for different orbits still look the same; in fact, they are the same.

Proposition 4.3.19 *Let $f\colon S^1 \to S^1$ be an orientation-preserving homeomorphism of S^1 without periodic points. Then $\omega(x)$ is independent of x and $E := \omega(x)$ is perfect and either S^1 or nowhere dense (see Definition A.1.5).*

By Proposition A.1.7, perfect nowhere dense sets are Cantor sets, that is, they are homeomorphic to the standard middle-third Cantor set. Therefore, this result produces Cantor sets directly from the dynamics of a circle map – at least when we give an example where this is the possibility that is actually realized.

Proof Independence of x: We need to show that $\omega(x) = \omega(y)$ for $x, y \in S^1$. Let $z \in \omega(x)$. Then there is a sequence l_n in $\mathbb{N}$ such that $f^{l_n}(x) \to z$. If $y \in S^1$, then by Lemma 4.3.16 there exist $k_m \in \mathbb{N}$ such that $f^{k_m}(y) \in I_m := [f^{l_m}(x), f^{l_{m+1}}(x)]$. But then $\lim_{m\to\infty} f^{k_m}(y) = z$ and thus $z \in \omega(y)$.

Therefore $\omega(x) \subset \omega(y)$ for all $x, y \in S^1$ and by symmetry $\omega(x) = \omega(y)$ for all $x, y \in S^1$.

$E := \omega(x)$ is either S^1 or nowhere dense: Let us first show that E is the smallest closed nonempty f-invariant set. If $\varnothing \neq A \subset S^1$ is closed and f-invariant and $x \in A$, then $\{f^k(x)\}_{k\in\mathbb{Z}} \subset A$ since A is invariant and $E = \omega(x) \subset \overline{\{f^k(x)\}_{k\in\mathbb{Z}}} \subset A$ since A is closed. Thus any closed invariant set A is either empty or contains E. In particular, $\varnothing$ and E are the only closed invariant subsets of E itself. Since E is closed, it contains its boundary, which is itself a closed set (Exercise 2.6.6). The boundary is also invariant because a boundary point is a point for any neighborhood U of which we have $U \cap E \neq \varnothing$ and $U \setminus E \neq \varnothing$, a property that persists when we apply a homeomorphism. Therefore the boundary ∂E of E is a closed invariant subset of E and as such we must have either $\partial E = \varnothing$ and hence $E = S^1$, or else $\partial E = E$, which implies that E is nowhere dense (Exercise 2.6.6).

It remains to show that E is perfect. Let $x \in E$. Since $E = \omega(x)$, there is a sequence k_n such that $\lim_{n\to\infty} f^{k_n}(x) = x$. Since there are no periodic orbits, $f^{k_n}(x) \neq x$ for all n. Consequently, x is an accumulation point of E since $f^{k_n}(x) \in E$ for all n by invariance. □

4.3.6 Comparison and Classification

Both in Proposition 4.3.12 and in Proposition 4.3.19 there is a set of distinguished orbits (either periodic or in E) to which all others are asymptotic. This distinguished set corresponds most closely to the rotation with the same rotation number (for irrational rotation number this becomes clear with Theorem 4.3.20). Thus if there are periodic points, there is a remnant of the rotation that may be as small as a single periodic orbit or a finite set of them; otherwise, the corresponding remnant is at least a Cantor set. It is in this distinction that Proposition 4.3.19 shows that the orbit structure of maps without periodic points is quite different from that of maps with periodic points. If there are periodic points, all orbits are either periodic or asymptotic to a periodic orbit; otherwise, either all orbits are dense or all orbits are asymptotic to or in a Cantor set. Moreover, a further difference appears when we compare the orbit structure of a circle map with that of a rotation with the same rotation number. The vast majority of circle maps with periodic points possess nonperiodic orbits – Proposition 4.4.10 and Lemma 4.4.12 show that having nonperiodic orbits occurs over entire parameter intervals in a family of maps, whereas having all orbits periodic happens only for an instant. (Furthermore, similar arguments show that even having infinitely many periodic points is unstable, and hence rare.) Thus, the presence of nonperiodic orbits, which is a qualitative difference to a rational rotation, is the most common behavior for maps with rational rotation number.

For irrational rotation number the picture is different. The greatest qualitative similarity to an irrational rotation occurs when $E = S^1$ in Proposition 4.3.19. In this case all orbits are dense ($\omega(x) = S^1$ for all $x \in S^1$), which is the same situation as for an irrational rotation. Unlike in the case of rational rotation number, there is no indication that the alternative situation (E is a Cantor set) occurs more frequently (in fact, it never happens at all for C^2 maps). Indeed, a map with irrational rotation number ρ is either equivalent to or "contains" R_ρ, up to some distortion, according to whether its orbits are dense:

Theorem 4.3.20 (Poincaré Classification Theorem) *Let $f\colon S^1 \to S^1$ be an orientation-preserving homeomorphism with irrational rotation number ρ. Then there is a continuous monotone map $h\colon S^1 \to S^1$ with $h \circ f = R_\rho \circ h$.*

(1) *If f is transitive, then h is a homeomorphism.*
(2) *If f is not transitive, then h is noninvertible.*

The map h here plays the role of the changes of variable or conjugacies that we encountered in Section 1.2.9.3 and Section 3.1.3, except that it may not be invertible. Section 4.3.3 rules out the nontransitive case for smooth f.

Proof We first construct the lift of h only on the lift of a single orbit and show that it is monotone. We then extend it to the closure of that lift and, using monotonicity, "fill in" any gaps that may be left. Finally we define h as the projection.

Pick a lift $F\colon \mathbb{R} \to \mathbb{R}$ of f and $x \in \mathbb{R}$. Let $B := \{F^n(x) + m\}_{n,m\in\mathbb{Z}}$ be the total lift of the orbit of $[x]$. Define $H\colon B \to \mathbb{R}$, $F^n(x) + m \mapsto n\rho + m$, where $\rho := \rho(F)$. By Proposition 4.3.15, this map is monotone, and $H(B)$ is dense in $\mathbb{R}$ by Proposition 4.1.1. If we write $\tilde{R}_\rho\colon \mathbb{R} \to \mathbb{R}$, $x \mapsto x + \rho$, then $H \circ F = \tilde{R}_\rho \circ H$ on B because

$$H \circ F(F^n(x) + m) = H(F^{n+1}(x) + m) = (n+1)\rho + m$$

and

$$\tilde{R}_\rho \circ H(F^n(x) + m) = \tilde{R}_\rho(n\rho + m) = (n+1)\rho + m.$$

Lemma 4.3.21 *H has a continuous extension to the closure $\bar{B}$ of B.*

Proof If $y \in \bar{B}$, then there is a sequence $(x_n)_{n\in\mathbb{N}}$ in B such that $y = \lim_{n\to\infty} x_n$. To show that $H(y) := \lim_{n\to\infty} H(x_n)$ exists and is independent of the choice of a sequence approximating y, observe first that the left and right limits exist and are independent of the sequence since H is monotone. If the left and right limits disagree, then $\mathbb{R} \setminus H(B)$ contains an interval, which contradicts the density of $H(B)$. □

H can now easily be extended to $\mathbb{R}$: Since $H\colon \bar{B} \to \mathbb{R}$ is monotone and surjective [because H is monotone and continuous on B, $\bar{B}$ is closed, and $H(B)$ is dense in $\mathbb{R}$] there is no choice in defining H on the intervals complementary to $\bar{B}$: Set $H =$ const. on those intervals, choosing the constant equal to the values at the endpoints. This gives a map $H\colon \mathbb{R} \to \mathbb{R}$ such that $H \circ F = \tilde{R}_\rho \circ H$ and thus the desired map $h\colon S^1 \to S^1$ since for $z \in B$ we have

$$H(z+1) = H(F^n(x) + m + 1) = n\rho + m + 1 = H(z) + 1,$$

and this property persists under continuous extension.

To decide invertibility note that in the transitive case we start from a dense orbit and so $\bar{B} = \mathbb{R}$ and h is a bijection. In the nontransitive case, H is constant on the intervals complementary to the orbit closure that we used. □

Remark 4.3.22 In the transitive case of Theorem 4.3.20, when h is invertible we say that h *conjugates* f to R_ρ; in the case of noninvertible h we say that R_ρ is a *factor* of f via h. These notions are explored in Chapter 7 (Definition 7.3.3).

EXERCISES

Exercise 4.3.1 For which values of a does the function $F(x) = 2x + a$ define the lift of a circle map?

Exercise 4.3.2 Referring to (4.1.1), prove that $\rho(R_\alpha) = [\alpha]$.

Exercise 4.3.3 Consider $F(x) := x + (1/2)\sin x$. Decide whether F is the lift of a circle homeomorphism.

Exercise 4.3.4 Consider $F(x) := x + (1/4\pi)\sin 2\pi x$. Decide whether F is the lift of a circle homeomorphism, and, if so, decide whether that homeomorphism is orientation-preserving. If it is, determine the rotation number.

Exercise 4.3.5 Let $f\colon S^1 \to S^1$ be a monotone (but not necessarily invertible) map of degree one, that is, its lift is a monotone function $F\colon \mathbb{R} \to \mathbb{R}$ such that $F(x+1) = F(x) + 1$. Prove that the assertions of Proposition 4.3.5, Proposition 4.3.8 and Proposition 4.3.9 hold for f.

Exercise 4.3.6 Referring to Proposition 4.3.9, what happens with the rotation number under an orientation-reversing conjugacy?

Exercise 4.3.7 Let $f\colon S^1 \to S^1$ be a continuous map of degree one (not necessarily monotone) and $F\colon \mathbb{R} \to \mathbb{R}$ its lift. Prove that

$$\rho^+(F) := \lim_{n\to\infty} \max_{x\in S^1} \frac{F^n(x) - x}{n} \quad \text{and} \quad \rho^-(F) := \lim_{n\to\infty} \min_{x\in S^1} \frac{F^n(x) - x}{n}$$

both exist.

PROBLEMS FOR FURTHER STUDY

Problem 4.3.8 Under the assumptions of the previous exercise call

$$R(F) := \left\{\rho \in \mathbb{R} \mid \exists x \in \mathbb{R} \ \lim_{n\to\infty} \frac{F^n(x) - x}{n} = \rho\right\}$$

the *rotation set* of F. Prove that $R(F) \neq \varnothing$.

Problem 4.3.9 Prove that a circle homeomorphism with finitely many fixed points and an attracting fixed point has a repelling fixed point.

Show that there exists a circle homeomorphism with an attracting fixed point and without repelling fixed points.

4.4 CANTOR PHENOMENA

In Proposition 4.3.19, a Cantor set appears naturally for some circle homeomorphisms without periodic points. There are several other ways in which Cantor sets and related structures appear in this context. The conjugacy above is a case in point in the nontransitive case. The dependence of the rotation number on a parameter is an example with interesting physical implications.

4.4.1 Devil's Staircases

In the nontransitive case the map h in Theorem 4.3.20 is necessarily an example of the following interesting phenomenon:

Definition 4.4.1 A monotone continuous function $\phi\colon [0,1] \to \mathbb{R}$ (or $\phi\colon [0,1] \to S^1$) is called a *devil's staircase* if there exists a family $\{I_\alpha\}_{\alpha\in A}$ of disjoint closed subintervals of [0,1] of nonzero length with dense union such that ϕ takes distinct constant values on these subintervals. (See Figure 4.4.1.)

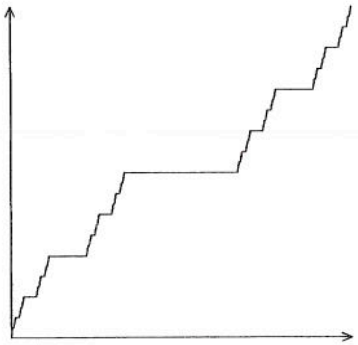


Figure 4.4.1. Devil's staircasse.

Example 4.4.2 A devil's staircase can be constructed in a fairly explicit way. For $x = 0.\alpha_1\alpha_2\alpha_3\cdots = \sum_{i=1}^{\infty}\alpha_i 3^{-i}$ ($\alpha_i \neq 1$) in the ternary Cantor set C define $f(x) := \sum_{i=1}^{\infty}\alpha_i 2^{-i-1} \in [0,1]$ as in Lemma 2.7.3. In Section 2.7.1 we found that f is surjective and nondecreasing, and that the two endpoints of a deleted interval are mapped to the same point. It is not hard to see that f is continuous (this is used in Problem 2.7.6). We can extend f to a nondecreasing continuous map on $[0,1]$ by defining it to be constant on each deleted interval, the constant being the common value of f at the endpoints. This is then a devil's staircase, also called a Cantor function.

The graph of this function has some self-similarity: The transformation given by $\left(\begin{smallmatrix} 1/3 & 0 \\ 0 & 1/2 \end{smallmatrix}\right)$ in the plane maps the graph to a proper subset of itself because $f(x) = f(3x)/2$ on $[0, 1/3]$.

The terminology "devil's staircase" refers to the odd situation that the graph of this function consists entirely of "steps", namely, the horizontal portions over the deleted intervals, yet there are no jumps at all; the function is continuous. Thus the tops of the steps are there, but not their "faces". In analysis this provides a quaint example with several odd properties, but we have now seen that in dynamics such functions come up rather naturally.

Let us revisit the construction of the map h above in order to understand the nontransitive case better. Since the set $\bar{B}$ from the proof of Theorem 4.3.20 projects to the closure of the orbit of $[x]$, it contains the ω-limit set $E = \omega([x])$ of $[x]$, and, by choosing $x \in \pi^{-1}(E)$, we obtain $\pi(\bar{B}) = E$, where E is the universal ω-limit set discussed previously. In the transitive case $\bar{B} = \mathbb{R}$ and $E = S^1$, but in the nontransitive case we find that if $x \in \pi^{-1}(E)$, then $\pi(\bar{B}) = E$ is a Cantor set. Consequently, in the nontransitive case h is a bijection of the identification space $E/\sim$ (identifying the two endpoints of each complementary interval) to S^1 and conjugates $f_{\restriction E/\sim}$ to $R_{\rho(f)}$. All orbits of f in E are dense in E (by the definition of E). On the other hand, the construction of $E = \omega(x)$ yields that all points outside E are attracted to E in both positive and negative time because iterates of such a point have to stay inside disjoint complementary intervals of E and the length of these goes to zero.

4.4.2 Wandering Domains

Conversely, one can think of the nontransitive map as being obtained from an irrational rotation by "blowing up" some orbits to intervals whose union then makes up the complement of E. These complementary intervals are thus permuted like the points on an orbit for an irrational rotation. All interior points in those intervals are "wandering" in the sense below since they stay within those intervals whose images are all disjoint. The next subsection has an explicit construction of such an example.

Definition 4.4.3 A point is said to be *wandering* if it has a neighborhood all of whose images and preimages are pairwise disjoint.

This behavior is the extreme opposite of recurrence, which we introduce in Definition 6.1.8.

To return to our comparison with the case of rational rotation number we note that in that case a map f is only conjugate to a rotation if all orbits are periodic with the same period and hence $f^q = \mathrm{Id}$ for some $q \in \mathbb{Z}$. Furthermore, a rational rotation can only be a factor when there are infinitely many periodic points, which, as we noted earlier, is unstable.

4.4.3 The Denjoy Example

We now give an example of a nontransitive circle diffeomorphism without periodic points. The construction starts with an irrational rotation and replaces the points of one orbit by suitably chosen intervals. The resulting map is not transitive. This example due to Arnaud Denjoy proves:

Proposition 4.4.4 *For $\rho \in \mathbb{R} \setminus \mathbb{Q}$ there is a nontransitive C^1 diffeomorphism $f\colon S^1 \to S^1$ with $\rho(f) = \rho$.*

Proof If $l_n := (|n|+3)^{-2}$ and $c_n := 2\left((l_{n+1}/l_n) - 1\right) \geq -1$, then

$$\sum_{n\in\mathbb{Z}} l_n < 2\sum_{n=0}^{\infty} l_n = 2\sum_{n=3}^{\infty}\frac{1}{n^2} < 2\int_2^{\infty}\frac{1}{x^2}\,dx = 1.$$

To "blow up" the orbit $x_n = R_\rho^n x$ of the irrational rotation R_ρ to intervals I_n of length l_n, insert the intervals I_n into S^1 so that they are ordered in the same way as the points x_n and the space between any two such intervals I_m and I_n is

$$\left(1 - \sum_{n\in\mathbb{Z}} l_n\right) d(x_m, x_n) + \sum_{x_k \in (x_m, x_n)} l_k\,.$$

(This is the sum of the lengths of the intervals I_k inserted in between and the length of the arc of the circle between x_m and x_n, appropriately scaled to reflect the fact that the total length of $S^1 \setminus \bigcup_{n\in\mathbb{Z}} I_n$ is $1 - \sum_{n\in\mathbb{Z}} l_n$.) To define a circle homeomorphism f such that $f(I_n) = I_{n+1}$ and $f_{\restriction S^1 \setminus \bigcup_{n\in\mathbb{Z}} I_n}$ is semiconjugate to a rotation it suffices to specify the derivative $f'(x)$ since f is then obtained by integration.

On the interval $[a, a+l]$ define the tent function

$$h(a, l, x) := 1 - \frac{1}{l}|2(x-a) - l|.$$

Then $h(a, l, a + l/2) = 1$ and $\int_a^{a+l} h(a, l, x)\,dx = l/2$. Denote the left endpoint of I_n by a_n and let

$$f'(x) = \begin{cases} 1 & \text{for } x \in S^1 \setminus \bigcup_{n\in\mathbb{Z}} I_n, \\ 1 + c_n h(a_n, l_n, x) & \text{for } x \in I_n. \end{cases}$$

The choice $c_n = 2((l_{n+1}/l_n) - 1) = 2(l_{n+1} - l_n)/l_n$ implies

$$\int_{I_n} f'(x)\,dx = \int_{I_n} (1 + c_n h(a_n, l_n, x))\,dx = l_n + \frac{l_n}{2} c_n = l_{n+1},$$

so indeed $f(I_n) = I_{n+1}$. □

Close inspection of this proof reveals that the derivative of the function f has to be somewhat distorted in order to contract intervals fast enough to fit into the interstices of the universal Cantor set. A systematic careful analysis shows that no sufficiently smooth circle homeomorphism exhibits this phenomenon.

A C^2 diffeomorphism $f\colon S^1 \to S^1$ with irrational rotation number $\rho(f) \in \mathbb{R} \setminus \mathbb{Q}$ is transitive and hence topologically conjugate to $R_{\rho(f)}$.

In fact, slightly weaker regularity hypotheses suffice. The most natural weakening is to assume merely that the derivative has bounded variation. A function $g\colon S^1 \to \mathbb{R}$ is said to be of *bounded variation* if its total variation $\mathrm{Var}(g) := \sup \sum_{k=1}^n |g(x_k) - g(x_k')|$ is finite. Here the sup is taken over all finite collections $\{x_k, x_k'\}_{k=1}^n$ such that x_k, x_k' are endpoints of an interval I_k and $I_k \cap I_j = \varnothing$ for $k \neq j$. Every Lipschitz function and hence every continuously differentiable function has bounded variation.

4.4.4 Dependence of the Rotation Number on a Parameter

Here we examine the dependence of the rotation number on the map as the map is varied. To begin with, it is continuous and monotone.

Proposition 4.4.5 *$\rho(\cdot)$ is continuous in the uniform topology.*

Proof If $\rho(f) = \rho$, take $p'/q', p/q \in \mathbb{Q}$ such that $p'/q' < \rho < p/q$. Pick the lift F of f for which $-1 < F^q(x) - x - p \leq 0$ for some $x \in \mathbb{R}$. Then $F^q(x) < x + p$ for all $x \in \mathbb{R}$, since otherwise $F^q(x) = x + p$ for some $x \in \mathbb{R}$ by the Intermediate-Value Theorem and $\rho = p/q$. Since the function $F^q - \mathrm{Id}$ is periodic and continuous, it attains its maximum. Thus there exists $\delta > 0$ such that $F^q(x) < x + p - \delta$ for all $x \in \mathbb{R}$. This implies that every sufficiently small perturbation $\bar{F}$ of F in the uniform topology also satisfies $\bar{F}^q(x) < x + p$ for all $x \in \mathbb{R}$ and thus $\rho(\bar{F}) < p/q$. A similar argument involving p'/q' completes the proof. □

The definition of the rotation number further suggests that it is monotone: If $F_1 > F_2$, then $\rho(F_1) \geq \rho(F_2)$ follows from the definition. This leads to the following concepts of ordering on the circle and for maps of the circle:

Definition 4.4.6 Define "$<$" on S^1 by $[x] < [y] :\Leftrightarrow y - x \in (0, 1/2) \pmod 1$ and define a partial ordering "$\prec$" on the collection of orientation-preserving circle homeomorphisms by $f_0 \prec f_1 :\Leftrightarrow f_0(x) < f_1(x)$ for all $x \in S^1$.

Notice that neither of these orderings is transitive. Indeed, $[0] < [1/3] < [2/3] < [0]$ and correspondingly $R_0 \prec R_{1/3} \prec R_{2/3} \prec R_0$, where R_α is the

rotation as in Section 4.1. The definition of rotation number now immediately implies:

Proposition 4.4.7 *$\rho(\cdot)$ is monotone: If $f_1 \prec f_2$, then $\rho(f_1) \leq \rho(f_2)$.*

Remark 4.4.8 In particular, if $\{f_t\}$ is a family of orientation-preserving circle homeomorphisms such that $f_t(x)$ is increasing in t for every $x \in \mathbb{R}$, then $\rho(f_t)$ is nondecreasing in t.

At irrational values the rotation number is strictly increasing:

Proposition 4.4.9 *If $f_0 \prec f_1$ and $\rho(f_0) \notin \mathbb{Q}$, then $\rho(f_0) < \rho(f_1)$.*

Proof If F_0 and F_1 are lifts with $0 < F_1(x) - F_0(x) < 1/2$ for all $x \in \mathbb{R}$, then by continuity and periodicity $F_1(x) - F_0(x) > \delta$ for some $\delta > 0$ and all $x \in \mathbb{R}$. Take $p/q \in \mathbb{Q}$ such that $p/q - \delta/q < \rho(F_0) < p/q$. Then there exists $x_0 \in \mathbb{R}$ such that $F_0^q(x_0) - x_0 > p - \delta$ [because otherwise $\rho(F_0) = \lim_{n\to\infty}(F_0^{nq}(x) - x/nq) \leq \lim_{n\to\infty}(n(p - \delta)/nq) = p/q - \delta/q$]. Since

$$\begin{aligned} F_1^q(x_0) = F_1\big(F_1^{q-1}(x_0)\big) &> F_0\big(F_1^{q-1}(x_0)\big) + \delta \\ &> F_0\big(F_0^{q-1}(x_0)\big) + \delta = F_0^q(x_0) + \delta > x_0 + p, \end{aligned}$$

we either have $F_1^q(x) > x + p$ for all $x \in \mathbb{R}$ or $F_1^q(x_1) = x_1 + p$ for some $x_1 \in \mathbb{R}$. In either case $\rho(F_0) < p/q \leq \rho(F_1)$. □

While Proposition 4.4.9 shows that having irrational rotation number is not stable, the situation is different for rational rotation number:

Proposition 4.4.10 *Let $f\colon S^1 \to S^1$ be an orientation-preserving homeomorphism with rational rotation number $\rho(f) = p/q$ and some nonperiodic points. Then all sufficiently nearby perturbations $\bar{f}$ with $\bar{f} \prec f$ or all sufficiently nearby perturbations $\bar{f}$ with $f \prec \bar{f}$ (or both) have rotation number p/q.*

The basic issue is whether the graph of $F^q - p$ has portions above and below the diagonal, in which case small perturbations either way cannot get rid of intersections with the diagonal (Figure 4.4.2). The borderline case, in which the graph lies entirely on one side, is exactly the one where the bifurcation to different dynamics occurs (see also Figure 2.3.2).

Proof Since f has nonperiodic points, $F^q - \mathrm{Id} - p$ does not vanish identically for any lift F of f. (It does have zeros by assumption.) If there exists $x \in \mathbb{R}$ with

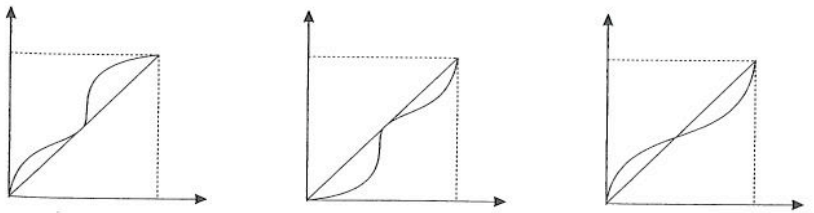


Figure 4.4.2. One-sided and two-sided stability.

$F^q(x) - x - p > 0$, then for any sufficiently small perturbation $\bar{f} \prec f$ the corresponding lift $\bar{F}$ of $\bar{f}$ is such that $\bar{F}^q(x) - x - p > 0$ and hence $\rho(\bar{f}) \geq p/q$; so $\rho(\bar{f}) = p/q$ by Proposition 4.4.7. Otherwise, the same holds for perturbations with $f \prec \bar{f}$. □

Remark 4.4.11 The proof shows that circle maps that have an attracting or repelling periodic orbit (an orbit that lifts to a point where $F^q - \mathrm{Id} - p$ changes sign) can be perturbed (in either direction) without changing the rotation number.

On the other hand, if $F^q - \mathrm{Id} - p$ does not change sign, for example, $F^q - \mathrm{Id} - p \geq 0$, then any perturbation $\bar{f}$ with $f \prec \bar{f}$ has rotation number $\rho(\bar{f}) > p/q$ since $\bar{F}^q - \mathrm{Id} - p \geq \delta > 0$. In this case the zeros of $F^q - \mathrm{Id} - p$ project to "parabolic" or *semistable* periodic orbits. These are orbits p that attract on one side and repel on the other side; that is, there is some open neighborhood U of p such that for all x in one component of $U \setminus \{p\}$ we have $\lim_{n\to\infty} d(f^n(x), f^n(p)) = 0$, and for all x in the other component $\lim_{n\to-\infty} d(f^n(x), f^n(p)) = 0$ (see Figure 4.3.3).

Here is an extreme case.

Lemma 4.4.12 *If all points of a map $f\colon S^1 \to S^1$ are periodic, then the rotation number is strictly increasing at f.*

To see that the rotation number depends on f in a nonsmooth way we reformulate these conclusions: The rotation number as a function of a parameter can (and usually will) be a devil's staircase (see Definition 4.4.1).

Proposition 4.4.13 *Suppose that $(f_t)_{t\in[0,1]}$ is a monotone continuous family of orientation-preserving circle homeomorphisms such that $\rho\colon t \mapsto \rho(f_t)$ is nonconstant and there exists a dense set $S \subset \mathbb{Q}$ such that, for each map f_t, either $\rho(f_t) \notin S$ or f_t has some nonperiodic points. Then ρ is a devil's staircase.*

Proof By Proposition 4.4.5, ρ is monotone and continuous. Together with Proposition 4.4.10, this also implies that $\rho^{-1}(S)$ is a disjoint union of closed intervals of positive length.

We need to show that $\rho^{-1}(S)$ is dense. Assume, by enlarging S if necessary, that whenever $\rho(f_t) = p/q \in \mathbb{Q} \setminus S$, f_t has only periodic points. Then Proposition 4.4.9 and Lemma 4.4.12 imply that ρ is strictly monotone at points $t \in \rho^{-1}([0,1] \setminus S)$. Thus for $t \in [0,1) \setminus \rho^{-1}(S)$ and $\epsilon > 0$ we have $\rho(t) \neq \rho(t+\epsilon)$, and hence by the density of S, the continuity of ρ, and the Intermediate Value-Theorem there exists a $t_1 \in \rho^{-1}(S) \cap [t, t+\epsilon]$. This proves density. □

In closing we remark that the results of this section depend on the monotonicity and continuity of f, but not on invertibility. Thus it suffices to assume that $f\colon S^1 \to S^1$ is a continuous order-preserving map of degree one, that is, its lift F is nondecreasing and $F(x+1) = F(x) + 1$ (Exercise 4.3.5). Such a map may take constant values on a finite or countable set of intervals.

4.4.5 Frequency Locking

The understanding gained in the preceding subsection about the dependence of the rotation number on a parameter also leads to insights about flows on the 2-torus, and in particular about some systems of differential equations arising in applications. The phenomenon that we are able to shed some light on now is that coupled oscillators tend to become *synchronized*, that is, their frequencies will coincide or be at least rationally related.

Where do we stand? The problems of flashing fireflies and circadian rhythms were introduced in Section 1.2.10 as situations that one might model as coupled oscillators. We can simplistically model those biological clocks as a harmonic oscillator or something close. Indeed, the harmonic oscillator is a good starting point, as we will see by linearization in Section 6.2.2.

Now, in Section 4.2.4 we found that two uncoupled harmonic oscillators produce, on a joint level set, a linear toral flow. This linear flow on $\mathbb{T}^2$ satisfies the differential equations

$$\dot{x}_1 = \omega_1$$
$$\dot{x}_2 = \omega_2.$$

To get an impression of the effects of coupling the two oscillators, modify the preceding differential equations by including "mixed" terms:

$$\begin{aligned} \dot{x}_1 &= \omega_1 + c_1 \sin 2\pi(x_2 - x_1) \\ \dot{x}_2 &= \omega_2 + c_2 \sin 2\pi(x_1 - x_2). \end{aligned} \tag{4.4.1}$$

This is not exactly the same as coupling the original second-order equations in Section 4.2.4, but it is a good way to get some insight.

A small detail here is the choice of sines to produce the mixed terms. This makes sense because both variables are only defined mod 1. The constants c_1 and c_2 indicate the strength of the coupling. When they are both zero, there is no coupling and we are back to a linear flow on the 2-torus. If they are positive, the right-hand side acts to increase the slower rate of change of the two ω's and to slow down the faster one, which could plausibly lead to synchronization.

In Section 4.2.3 we learned to study flows on the 2-torus by looking at the section map of the flow of (4.4.1) for the section $x_2 = 0$, say. In the absence of the coupling terms, that is, when $c_1 = c_2 = 0$, this section map is just the rotation with rotation number ω_1/ω_2. For small values of the coupling constants the section map is therefore a perturbation of this rotation. In "most" cases this perturbation has a rational rotation number, because this is the stable situation. And whenever the rotation number is rational all asymptotic behavior is periodic (with the same period).

To explore this a little more carefully, suppose that ω_1 and ω_2 are close to each other. In fact, assume first that $\omega_1 = \omega_2 =: \omega$. In that case, $x(t) = y(t) = \omega t$ is a solution of (4.4.1). This particular solution works for all values of c_1 and c_2, so the section map always has a fixed point, and hence rotation number 0.

For $(c_1, c_2) \neq (0,0)$, the section maps are not conjugate to rotations and therefore their rotation number persists under small perturbations by Proposition 4.4.10. In particular, when we fix c_1 and c_2, then for small values of $\omega_1 - \omega_2$ the flow of (4.4.1) has a section map with a fixed point, all of whose orbits are asymptotic to a

CHAPTER 7

Simple Systems with Complicated Orbit Structure

This chapter presents a rich array of properties of a collection of examples. Its coherence derives from the fact that it is part of a general theory we outline in Chapter 10. The examples (other than the quadratic map f_4) are instances of hyperbolic dynamical systems (or symbolic dynamical systems), and the properties we derive here are largely properties common to hyperbolic and symbolic dynamical systems.

7.1 GROWTH OF PERIODIC POINTS

Periodic orbits represent the most distinctive special class of orbits. So far we have mostly encountered maps with few periodic orbits or, as in the case of a rational rotation, only periodic orbits. In these basic examples different periods did not appear for the same map. Even the most complex situations so far still involve periodic orbits neatly organized by period in families such as invariant curves in plane rotations, linear twists, the time-1 map for the mathematical pendulum, or billiards. There we placed more emphasis on coherence than complexity. Now we encounter the first examples with a different periodic pattern. In these cases, when periodic points of different periods are present, we want to count them.

Definition 7.1.1 For a map $f\colon X \to X$, let $P_n(f)$ be the number of periodic points of f with (not necessarily minimal) period n, that is, the number of fixed points for f^n.

This section introduces numerous new examples of dynamical systems. For now they are introduced with a view to their periodic orbit structure, but in due time numerous other fascinating features of their orbit structure will emerge.

7.1.1 Linear Expanding Maps

Consider the noninvertible map E_2 of the circle given in multiplicative notation by

$$E_2(z) = z^2, \qquad |z| = 1,$$

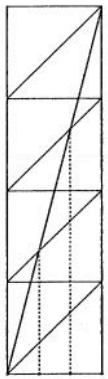


Figure 7.1.1. Periodic points for an expanding map.

and in additive notation by

$$E_2(x) = 2x \pmod 1. \tag{7.1.1}$$

Proposition 7.1.2 *$P_n(E_2) = 2^n - 1$ and periodic points for E_2 are dense in S^1.*

Proof If $E_2^n(z) = z$, then $z^{2^n} = z$, and $z^{2^n-1} = 1$. Thus every root of unity of order $2^n - 1$ is a periodic point for E_2 of period n. There are exactly $2^n - 1$ of these, and they are uniformly spread over the circle with equal intervals. In particular, when n becomes large these intervals become small. (See Figure 7.1.1) □

We see from Proposition 7.1.2 that a natural measure of asymptotic growth of the number of periodic points is the exponential growth rate $p(f)$ for the sequence $P_n(f)$:

$$p(f) = \overline{\lim_{n\to\infty}} \frac{\log_+ P_n(f)}{n}, \tag{7.1.2}$$

where $\log_+(x) = \log(x)$ for $x \geq 1$, 0 otherwise. In particular, Proposition 7.1.2 shows that $p(E_2) = \overline{\lim}_{n\to\infty}(\log 2^n + \log(1 - 2^{-n}))/n = \log 2$.

The maps

$$E_m\colon x \mapsto mx \pmod 1,$$

where m is an integer of absolute value greater than one, represent a straightforward generalization of the map E_2. Not surprisingly, these maps also have dense sets of periodic orbits. The proof of Proposition 7.1.2 holds verbatim with the replacement of 2 by m:

Proposition 7.1.3 *$P_n(E_m) = |m^n - 1|$ and periodic points for E_m are dense.*

Proof $z = E_m^n(z) = z^{m^n}$ has $|m^n - 1|$ solutions that are evenly spaced. □

See also Section 7.1.3.

Another property of the maps E_m worth noticing is preservation of length similar to the property of preservation of phase volume discussed in Section 6.1.

Naturally, the length of an image of any arc increases; however, if one considers the *complete preimage* of an arc Δ under E_m, one immediately sees that it consists of $|m|$ arcs of length $l(\Delta)/|m|$ each, placed along the circle at equal distances. The analysis in Section 6.1.2 can be extended to noninvertible volume-preserving maps, so recurrent points are dense in this situation as well.

7.1.2 Quadratic and Quadratic-Like Maps

For $\lambda \in \mathbb{R}$, let $f_\lambda : \mathbb{R} \to \mathbb{R}$, $f_\lambda(x) := \lambda x(1-x)$. For $0 \le \lambda \le 4$, the f_λ map the unit interval $I = [0,1]$ into itself. The family f_λ is referred to as the *quadratic family*. For $\lambda \le 3$, this family was discussed in detail in Section 2.5, and the asymptotic behavior for any such λ is fairly simple and changes with λ only a few times. As it turns out, for the remaining interval of parameter values the quadratic family develops a bewildering array of complicated but different types of behavior, which change with caleidoscopic speed (see Figure 7.1.2 and Chapter 11). Note that $P_n(f_\lambda) \le 2^n$ because the nth iterate of f_λ is a polynomial of degree 2^n, and hence the equation $(f_\lambda)^n(x) = x$ has at most 2^n solutions. While one may expect that in the complex plane this equation would indeed have exactly 2^n solutions for most values of the parameter λ, this is certainly not the case for real solutions.

Here we consider the behavior of the quadratic family for large values of the parameter, namely, $\lambda \ge 4$. While for $\lambda > 4$ the interval $[0,1]$ is not preserved, the set of points that remains in that interval is still quite interesting.

The analysis of the behavior of the quadratic family on the unit interval for $0 \le \lambda \le 3$ carried out in Section 2.5 showed simple periodic patterns: Only points of periods 1 and 2 appear, and their number is small. With moderate effort this analysis can be extended as far as $\lambda = 1 + \sqrt{6}$ (Proposition 11.2.1). On the other hand, we have:

Proposition 7.1.4 *For $\lambda \ge 4$ we have $P_n(f_\lambda) = 2^n$.*

Proof Since $P_n(f_\lambda) \le 2^n$, it suffices to prove the reverse inequality. To that end we use the following observation: If $f : \mathbb{R} \to \mathbb{R}$ is continuous and $\Delta \subset [0,1]$ is an interval such that one endpoint is mapped to 0 and the other to 1, then by the Intermediate-Value Theorem there is a fixed point of f in Δ. Now $[0,1] \subset [f_\lambda(0), f_\lambda(1/2)]$ and $[0,1] \subset [f_\lambda(1/2), f_\lambda(1)]$, so there are intervals $\Delta_0 \subset [0,1/2]$ and $\Delta_1 \subset [1/2,1]$ whose images under f_λ are exactly $[0,1]$, giving us two fixed points for f. The nonzero fixed

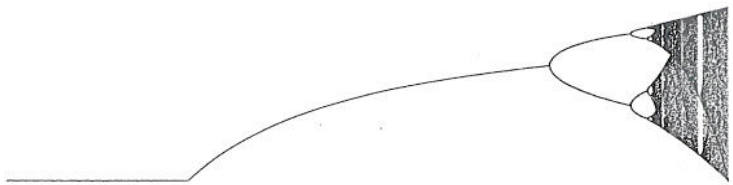


Figure 7.1.2. Bifurcation diagram.

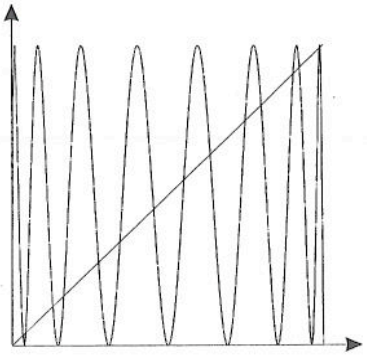


Figure 7.1.3. Periodic points of f_4.

point is indeed in the interior of Δ_1 because the right endpoint of Δ_1 is 1 and hence is mapped to 0, so the other endpoint is mapped to 1 and therefore neither are fixed.

Furthermore, the preimages of Δ_0 and Δ_1 under f consist of two intervals each, so there are four intervals whose images under f^2 are exactly $[0,1]$. Each contains a fixed point of f_λ^2, again every one except 0 being in the interior of the corresponding interval, so no two of these fixed points coincide.

Repeating this argument successively for higher iterates of f_λ we obtain 2^n intervals whose images under f_λ^n are $[0,1]$, and each of which therefore contains at least one fixed point, giving 2^n distinct orbits of period n for f_λ. □

It is useful that the argument to show that $P_n(f_\lambda) \ge 2^n$ applies to any continuous map $f : [0,1] \to \mathbb{R}$ with $f(0) = f(1) = 0$ and such that there is a $c \in [0,1]$ with $f(c) \ge 1$. In this more general case it is somewhat more convenient, however, to talk about intervals whose images under f^n contain $[0,1]$ rather than being exactly $[0,1]$.

In the quadratic case (for $\lambda > 4$) one can refine the preceding argument slightly to show that there are exactly 2^n periodic points (rather than using that the degree of f^n is 2^n). This also works for some continuous maps f of this more general nature, which are monotone on $[0,c]$ as well as $[c,1]$. A continuous map defined on an interval that is increasing to the left of an interior point and decreasing thereafter is said to be *unimodal*. Thus we have found

Proposition 7.1.5 *If $f : [0,1] \to \mathbb{R}$ is continuous, $f(0) = f(1) = 0$, and there exists $c \in [0,1]$ such that $f(c) > 1$, then $P_n(f) \ge 2^n$. If, in addition, f is unimodal and expanding (that is, $|f(x) - f(y)| > |x - y|$) on each interval of $f^{-1}((0,1))$, then $P_n(f) = 2^n$.*

The heart of the proof is the following lemma:

Lemma 7.1.6 *Denote by $\mathcal{M}_k$ the collection of continuous maps $f : [0,1] \to \mathbb{R}$ such that $f^{-1}((0,1)) = \bigcup_{i=1}^k I_i$ with $I_i \subset [0,1]$ open intervals, f monotonic on I_i, and $f(I_i) = (0,1)$. Then $f \circ g \in \mathcal{M}_{kl}$ whenever $f \in \mathcal{M}_k$ and $g \in \mathcal{M}_l$.*

Proof If $f \in \mathcal{M}_k$ and $g \in \mathcal{M}_l$, then $f^{-1}((0,1)) = \bigcup_{i=1}^{k} I_i$ and $g^{-1}(I_i) = \bigcup_{j=1}^{l} J_{ij}$ with $\{J_{ij} \mid 1 \le i \le k,\ 1 \le j \le l\}$ pairwise disjoint and $(f \circ g)^{-1}((0,1)) = \bigcup_{ij} J_{ij}$. The composition $f \circ g$ is monotonic on J_{ij} and $f \circ g(J_{ij}) = (0,1)$. □

Proof of Proposition 7.1.5 The lemma shows that $P_n(f) \ge k^n$ for $f \in \mathcal{M}_k$ because $f^n \in \mathcal{M}_{k^n}$. If f is expanding on every interval of $f^{-1}((0,1))$, then the same holds for iterates of f. This shows that on each of those intervals there is at most one solution of $f^n(x) = x$. Therefore, $P_n(f) \le k^n$, proving equality. □

7.1.3 Expanding Maps and Degree

Next we consider a nonlinear generalization of the expanding maps E_m. We use additive notation for circle maps. In this notation derivatives of maps can be expressed as real-valued functions.

Definition 7.1.7 A continuously differentiable map $f\colon S^1 \to S^1$ is said to be an *expanding* map if $|f'(x)| > 1$ for all $x \in S^1$.

Since f' is continuous and periodic, the minimum of $|f'|$ is attained and hence is greater than 1.

Proposition 4.3.1 gives us a function $F\colon \mathbb{R} \to \mathbb{R}$ that satisfies $[F(x)] = f([x])$ and $F(s+1) = F(s) + \deg(f)$, where $\deg(f)$ is the *degree* of f. It has the following simple property:

Lemma 7.1.8 *If $f, g\colon S^1 \to S^1$ are continuous, then* $\deg(g \circ f) = \deg(f)\deg(g)$, *in particular* $\deg(f^n) = \deg(f)^n$.

Proof If F, G are lifts of f and g, respectively, then $G(s+k) = G(s+k-1) + \deg(g) = \cdots = G(s) + k\deg(g)$ and $G(F(s+1)) = G(F(s) + \deg(f)) = G(F(s)) + \deg(g)\deg(f)$. □

This property is useful for counting periodic points.

Proposition 7.1.9 *If $f\colon S^1 \to S^1$ is an expanding map, then* $|\deg(f)| > 1$ *and* $P_n(f) = |\deg(f)^n - 1|$.

Proof $|f'| > 1$ implies $|F'| > 1$ for any lift, so, by the Mean-Value Theorem A.2.3, $|\deg(f)| = |F(x+1) - F(x)| > 1$. By the chain rule an iterate of an expanding map is itself expanding, so by Lemma 7.1.8 it suffices to consider the case $n = 1$. Take a lift F of f and consider it on the interval $[0,1]$. The fixed points of f are the projections of the points x for which $F(x) - x \in \mathbb{Z}$. The function $g(x) := F(x) - x$ satisfies $g(1) = g(0) + \deg(f) - 1$, so by the Intermediate-Value Theorem there are at least $|\deg(f) - 1|$ points x where $g(x) \in \mathbb{Z}$. If $g(0) \in \mathbb{Z}$, then there are $|\deg(f) - 1| + 1$ such points, but 0 and 1 project to the same point on S^1. Now $g'(x) \ne 0$, so g is strictly monotone and hence takes every value at most once. Thus there are exactly $|\deg(f) - 1|$ fixed points on S^1. □

This proposition in particular establishes part of the analog of Proposition 7.1.2 for E_m.

Similarly to quadratic maps, the argument that shows $P_n(f) \ge |\deg(f)^n - 1|$ works for any continuous map. It is trivial for maps of degree 1 because the assertion is vacuous. Indeed, irrational rotations do not have any fixed or periodic points. For maps of degree 0 it merely guarantees a fixed point. For maps f with $|\deg(f)| > 1$, however, this result gives exponential growth of the number of periodic points: $p(f) \ge \log_+(|\deg(f)|)$.

7.1.4 Hyperbolic Linear Map of the Torus

The previous examples were all one-dimensional, but the patterns of the growth and distribution of periodic points observed in those examples also appear in higher dimension.

A convenient model to demonstrate this is built from the following linear map of $\mathbb{R}^2$:

$$L(x,y) = (2x+y, x+y) = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

If two vectors (x,y) and (x',y') represent the same element of $\mathbb{T}^2$, that is, if $(x-x', y-y') \in \mathbb{Z}^2$, then $L(x,y) - L(x',y') \in \mathbb{Z}^2$, so $L(x,y)$ and $L(x',y')$ also represent the same element of $\mathbb{T}^2$. Thus L defines a map $F_L\colon \mathbb{T}^2 \to \mathbb{T}^2$:

$$F_L(x,y) = (2x+y, x+y) \pmod 1.$$

The map F_L is, in fact, an automorphism of the torus viewed as an additive group. It is invertible because the matrix $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ has determinant one, so L^{-1} also has integer entries [in fact $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)^{-1} = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 2 \end{smallmatrix}\right)$] and hence defines a map $F_{L^{-1}} = F_L^{-1}$ on $\mathbb{T}^2$ by the same argument. The eigenvalues of L are

$$\lambda_1 = \frac{3+\sqrt{5}}{2} > 1 \quad \text{and} \quad \lambda_1^{-1} = \lambda_2 = \frac{3-\sqrt{5}}{2} < 1. \tag{7.1.3}$$

Figure 7.1.4 gives an idea of the action of F_L on the fundamental square $I = \{(x,y) \mid 0 \le x < 1,\ 0 \le y \le 1\}$. The lines with arrows are the eigendirections. For any matrix L with determinant ± 1, the map F_L preserves the area of sets on the torus.

Proposition 7.1.10 *Periodic points of F_L are dense and* $P_n(F_L) = \lambda_1^n + \lambda_1^{-n} - 2$.

Proof To obtain density we show that points with rational coordinates are periodic points. Let $x, y \in \mathbb{Q}$. Taking the common denominator write $x = s/q$, $y = t/q$, where $s, t, q \in \mathbb{Z}$. Then $F_L(s/q, t/q) = ((2s+t)/q, (s+t)/q)$ is a rational point whose coordinates also have denominator q. But there are only q^2 different points on $\mathbb{T}^2$ whose coordinates can be represented as rational numbers with denominator q, and all iterates $F_L^n(s/q, t/q)$, $n = 0, 1, 2\ldots$, belong to that finite set. Thus they must repeat, that is, $F_L^n(s/q, t/q) = F_L^m(s/q, t/q)$ for some $n, m \in \mathbb{Z}$. But since F_L is invertible, $F_L^{n-m}(s/q, t/q) = (s/q, t/q)$ and $(s/q, t/q)$ is a periodic point, as required. This gives density. (By contrast, not all rational points are periodic for E_m. See Exercise 7.1.1.)

Next we show that points with rational coordinates are the only periodic points for F_L. Write $F_L^n(x,y) = (ax+by,\ cx+dy) \pmod 1$, where $a, b, c, d \in \mathbb{Z}$. If

up ad infinitum. This is analogous to the construction of the ternary Cantor set, where an interval becomes two, then four, and so on.

The definitive geometric realization is carried out in Section 13.2 and illustrated in Figure 13.2.1 and on the cover of this book. This picture is representative of a great wealth of ideas in dynamics and deserves to be an icon for chaotic dynamics. Together with the horseshoe and linear toral automorphisms, the expanding map E_2 and the solenoid are the most tractable representatives of hyperbolic dynamical systems, and these have provided the framework of concepts and techniques within which each chaotic dynamical system is studied and described. This framework is developed in this chapter and the next, and it is described further in Chapter 10.

EXERCISES

Exercise 7.1.1 Prove that for the expanding map $E_m(|m| \geq 2)$ rational points are preimages of periodic points ("eventually periodic").

Exercise 7.1.2 Find a necessary and sufficient condition for a rational point to be periodic under E_m.

Exercise 7.1.3 Carry out the proof of Proposition 7.1.3 for the case $m < -1$.

Exercise 7.1.4 Prove that for any $n \in \mathbb{N}$ and $\lambda \geq 4$ the quadratic map f_λ has a periodic point whose *minimal* period (Definition 2.2.6) is n.

Exercise 7.1.5 Give an example of a continuous map $f\colon [0,1] \to \mathbb{R}$ with $f(0) = f(1) = 0$ for which there exists $c \in [0,1]$ such that $f(c) > 1$, and such that $P_n(f) > 2^n$.

Exercise 7.1.6 Give an example of a smooth unimodal map f such that $P_n(f) < 2^n$.

Exercise 7.1.7 Show that a continuous map f of S^1 can be deformed to $E_{\deg(f)}$, that is, that there is a continuous map $F\colon [0,1] \times S^1 \to S^1$ with $F(0,\cdot) = E_{\deg(f)}$ and $F(1,\cdot) = f$.

Exercise 7.1.8 Show that maps of different degrees cannot be deformed into each other, that is, that there is no continuous map $F\colon [0,1] \times S^1 \to S^1$ such that $\deg(F(0,\cdot)) \neq \deg(F(1,\cdot))$.

Exercise 7.1.9 Suppose $f\colon S^1 \to S^1$ has degree 2 and 0 is an attracting fixed point. Show that $P_n(f) > 2^n$.

Exercise 7.1.10 Consider the Fibonacci sequence from Section 1.2.2, Example 2.2.9, and Section 3.1.9. Show that the sequence obtained from taking the last digit of each Fibonacci number is periodic.

Exercise 7.1.11 Apply the inverse limit construction to a homeomorphism and prove that the result is naturally equivalent to the original system.

PROBLEMS FOR FURTHER STUDY

Problem 7.1.12 Prove that the solenoid in Section 7.1.5 is connected but not path-connected.

7.2 TOPOLOGICAL TRANSITIVITY AND CHAOS

We will show that some of the examples considered in the previous section are topologically transitive in the sense of Definition 4.1.3, that is, they have dense orbits. That there are at the same time infinitely many periodic points makes these examples different from irrational rotations and the other topologically transitive examples of Chapter 4 and Chapter 5. In expanding maps and hyperbolic linear maps of the torus we even found that the union of the periodic points is dense, which means that dense and periodic orbits are inextricably intertwined.

Thus, the global orbit structure is far more complex in these examples. This intertwining of density and periodicity is an essential feature of the complexity of the orbit structure. It causes sensitive dependence of any orbit on its initial conditions (see Definition 7.2.11 and Theorem 7.2.12), which is regarded as an essential ingredient of *chaos*.

Definition 7.2.1 A continuous map $f\colon X \to X$ of a metric space is said to be *chaotic* if it is topologically transitive and its periodic points are dense.[1]

Circle rotations show that neither condition alone gives much complexity.

We will show presently that expanding and hyperbolic maps are chaotic. In fact, we show the stronger property of topological mixing (Definition 7.2.5), which is absent in the minimal examples of Chapter 4 and Chapter 5. Before introducing the mixing property, we give an alternative definition of topological transitivity.

7.2.1 A Criterion for Topological Transitivity

We defined topological transitivity as the existence of a dense orbit. However, it is useful to have an alternate characterization in terms of subsets of phase space. In order to include noninvertible maps, we say that a sequence $(x_i)_{i\in\mathbb{Z}}$ is an orbit of f if $f(x_i) = x_{i+1}$ for all $i \in \mathbb{Z}$. However, we simply write $f^i(x)$ for $i \in \mathbb{Z}$ anyway to keep the notations more familiar.

Proposition 7.2.2 *Let X be a complete separable (that is, there is a countable dense subset) metric space with no isolated points. If $f\colon X \to X$ is a continuous map, then the following four conditions are equivalent:*

(1) *f is topologically transitive, that is, it has a dense orbit.*
(2) *f has a dense positive semiorbit.*
(3) *If $\varnothing \neq U, V \subset X$, then there exists an $N \in \mathbb{Z}$ such that $f^N(U) \cap V \neq \varnothing$.*
(4) *If $\varnothing \neq U, V \subset X$, then there exists an $N \in \mathbb{N}$ such that $f^N(U) \cap V \neq \varnothing$.*

Of course, the implications (4) $\Rightarrow$ (3) and (2) $\Rightarrow$ (1) are clear. To show which hypotheses are needed for which of the remaining directions, we prove Proposition 7.2.2 in the following form.

[1] There is no universally accepted definition of chaos, but this definition is equivalent to the one most commonly found in expository literature, which was put forward by Robert Devaney.

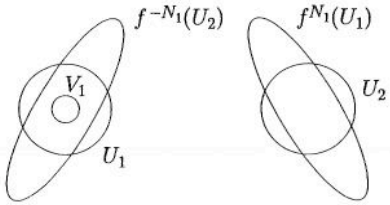


Figure 7.2.1. Construction for the proof.

Lemma 7.2.3 *Let X be a metric space and $f\colon X \to X$ a continuous map. Then* (1) *implies* (3). *If X has no isolated points, then* (1) *implies* (4). *If X is separable, then* (3) *implies* (1) *and* (4) *implies* (2).

Proof Let f be topologically transitive and suppose the orbit of $x \in X$ is dense. Then there exists an $n \in \mathbb{Z}$ such that $f^n(x) \in U$, and there is an $m \in \mathbb{Z}$ such that $f^m(x) \in V$; hence $f^{m-n}(U) \cap V \neq \varnothing$. This implies (3).

If we can choose $m > n$, then by taking $N := m - n$ we have even established (4). Otherwise we use the assumption that X has no isolated points, so $f^m(x)$ is not an isolated point and therefore there are $n_k \in \mathbb{Z}$ such that $|n_k| \to \infty$, $f^{n_k}(x) \in V$, and $f^{n_k}(x) \to f^m(x)$ as $k \to \infty$. Indeed, $n_k \to -\infty$ since $n_k \leq n$ by assumption (otherwise we are in the first case), so we can choose an $m' < 2m - n$ from among the n_k such that $f^{m'}(x) \in f^{m-n}(U)$. Then $x' := f^{n-m}(f^{m'}(x)) \in U$ and $f^{2m-n-m'}(x') = f^m(x) \in V$, so $f^N(U) \cap V \neq \varnothing$ with $N := 2m - n - m' \in \mathbb{N}$. Thus (1) $\Rightarrow$ (4) if X has no isolated points.

Now assume separability and one of the intersection conditions (3) and (4). We give one argument to prove both that (3) implies (1) and (4) implies (2). For a countable dense subset $S \subset X$, let $U_1, U_2, \ldots$ be the countable collection of balls centered at points of S with rational radius. We need to construct an orbit or semiorbit that intersects every U_n. By (3) there exists $N_1 \in \mathbb{Z}$ such that $f^{N_1}(U_1) \cap U_2 \neq \varnothing$. If (4) holds, we can take $N_1 \in \mathbb{N}$. Let V_1 be an open ball of radius at most $1/2$ such that $\bar{V}_1 \subset U_1 \cap f^{-N_1}(U_2)$. (See Figure 7.2.1.) There exists $N_2 \in \mathbb{Z}$ such that $f^{N_2}(V_1) \cap U_3$ is nonempty, and, if (4) holds, we can take $N_2 \in \mathbb{N}$. Again, take an open ball V_2 of radius at most $1/4$ such that $\bar{V}_2 \subset V_1 \cap f^{-N_2}(U_3)$. By induction, we construct a nested sequence of open balls V_n of radii at most 2^{-n} such that $\bar{V}_{n+1} \subset V_n \cap f^{-N_{n+1}}(U_{n+2})$. The centers of these balls form a Cauchy sequence whose limit x is the unique point in the intersection $V = \bigcap_{n=1}^{\infty} \bar{V}_n = \bigcap_{n=1}^{\infty} V_n$. Then $f^{N_{n-1}}(x) \in U_n$ for every $n \in \mathbb{N}$, and all $N_n \in \mathbb{N}$ if (4) holds.

If f is noninvertible, the last step may involve choices for negative values of N_n: Take i_k such that $N_{i_k} < 0$ for all k and $N_{i_{k+1}} < N_{i_k}$. Choose $x_0 = x$ and $x_{N_{i_k}} \in U_{i_k+1}$. Together with $f(x_k) = x_{k+1}$, this defines an orbit of x. □

Corollary 7.2.4 *A continuous open (Definition A.1.16) map f of a complete separable metric space without isolated points is topologically transitive if and only if there are no two disjoint open nonempty f-invariant sets.*

Proof If $U, V \subset X$ are open, then the invariant sets $\tilde{U} := \bigcup_{n\in\mathbb{Z}} f^n(U)$ and $\tilde{V} := \bigcup_{n\in\mathbb{Z}} f^n(V)$ are open because f is an open map, and therefore not disjoint by assumption, so $f^n(U) \cap f^m(V) \neq \varnothing$ for some $n, m \in \mathbb{Z}$. Then $f^{n-m}(U) \cap V \neq \varnothing$ and f is topologically transitive by Proposition 7.2.2. The other direction is obvious: A dense orbit visits every open set. □

7.2.2 Topological Mixing

There is a property of a dynamical system that immediately implies this criterion but is indeed much stronger:

Definition 7.2.5 A continuous map $f\colon X \to X$ is said to be *topologically mixing* if for any two nonempty open sets $U, V \subset X$ there is an $N \in \mathbb{N}$ such that $f^n(U) \cap V \neq \varnothing$ for every $n > N$.

By Proposition 7.2.2, every topologically mixing map is topologically transitive. On the other hand, our simple examples are not mixing. No translation T_γ, in particular no circle rotation, is topologically mixing. This follows from the fact that translations preserve the natural metric on the torus induced by the standard Euclidean metric on $\mathbb{R}^n$ and from the following general criterion.

Lemma 7.2.6 *Isometries are not topologically mixing.*

Proof Let $f\colon X \to X$ be an isometry (that is, a map that preserves the metric on X). Take distinct points $x, y, z \in X$, and let $\delta := \min(d(x, y), d(y, z), d(z, x))/4$. Let U, V_1, V_2 be δ-balls around x, y, z correspondingly. Since f preserves the diameter of any set, the diameter of $f^n(U)$ is at most 2δ whereas the distance between any $p \in V_1$ and $q \in V_2$ is greater than 2δ. Thus for each n either $f^n(U) \cap V_1 = \varnothing$ or $f^n(U) \cap V_2 = \varnothing$. □

7.2.3 Expanding Maps

For expanding maps we prove topological mixing by showing the stronger fact that, for any open set, its image under some iterate of the map contains S^1. For the linear expanding maps E_m this is obvious: Every open set contains an interval of the form $[l/|m|^k, (l+1)/|m|^k]$ for some integers k and $l \leq |m|^k$. The image of this interval under E_m^k is S^1.

Proposition 7.2.7 *Expanding maps of S^1 are topologically mixing.*

Proof Let $f\colon S^1 \to S^1$ such that $|f'(x)| \geq \lambda > 1$ for all x. Consider a lift F of f to $\mathbb{R}$. Then $|F'(x)| \geq \lambda$ for $x \in \mathbb{R}$. If $[a, b] \subset \mathbb{R}$ is an interval, then by the Mean-Value Theorem A.2.3 there exists a $c \in (a, b)$ such that $|F(b) - F(a)| = |F'(c)(b - a)| \geq \lambda(b - a)$ and so the length of any interval is increased by a factor at least λ^n under F^n. Consequently, for every interval I there exists $n \in \mathbb{N}$ such that the length of $F(I)$ exceeds 1. Thus the image of the projection of I to S^1 under f^n contains S^1. Since every open set of S^1 contains an interval, this shows that every open set has an image under an iterate of f that contains S^1. □

Corollary 7.2.8 *Linear expanding maps of S^1 are chaotic.*

Proof Transitivity follows from Proposition 7.2.7 and the density of periodic points from Proposition 7.1.3. □

For nonlinear expanding maps, this result also holds by invoking Theorem 7.4.3 (which is only stated for degree 2 in the following, but holds for any expanding map).

7.2.4 Hyperbolic Linear Map on the Torus

The hyperbolic linear map F_L of the torus induced by the linear map L with matrix $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ was introduced in Section 7.1.4. The eigenvectors corresponding to the first eigenvalue belong to the line $y = (\sqrt{5} - 1/2)x$. The family of lines parallel to it is invariant under L, and L uniformly expands distances on those lines by a factor λ_1. Similarly, there is an invariant family of contracting lines $y = (-\sqrt{5} - 1/2)x + \text{const}$.

Proposition 7.2.9 *The automorphism F_L is topologically mixing.*

Proof Fix open sets $U, V \subset \mathbb{T}^2$. The L-invariant family of lines

$$y = \frac{\sqrt{5}-1}{2}x + \text{const.} \tag{7.2.1}$$

projects to $\mathbb{T}^2$ as an F_L-invariant family of orbits of the linear flow T^t_ω with irrational slope $\omega = (1, (\sqrt{5} - 1)/2)$. By Proposition 5.1.3, this flow is minimal. Thus the projection of each line is everywhere dense on the torus, and hence U contains a piece J of an expanding line; furthermore, for any $\epsilon > 0$, there exists $T = T(\epsilon)$ and a segment of an expanding line of length T that intersects any ϵ-ball on the torus. Since all segments of a given length are translations of one another, this property holds for all segments. Now take ϵ such that V contains an ϵ-ball and $N \in \mathbb{N}$ such that $f^N(J)$ has length at least T. Then $f^n(J) \cap V \neq \varnothing$ for $n \geq N$ and thus $f^n(U) \cap V \neq \varnothing$ for $n \geq N$. □

Corollary 7.2.10 *The automorphism F_L is chaotic.*

Proof Combine Proposition 7.2.9, and Proposition 7.1.10. □

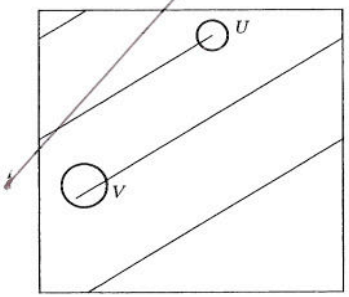


Figure 7.2.2. Topological mixing.

7.2.5 Chaos

At the outset of this section we motivated our definition of a chaotic map by saying that it implies sensitive dependence on initial conditions. We now justify this claim by defining and verifying sensitive dependence.

Definition 7.2.11 A map $f\colon X \to X$ of a metric space is said to exhibit *sensitive dependence* on initial conditions if there is a $\Delta > 0$, called a sensitivity constant, such that for every $x \in X$ and $\epsilon > 0$ there exists a point $y \in X$ with $d(x, y) < \epsilon$ and $d(f^N(x), f^N(y)) \geq \Delta$ for some $N \in \mathbb{N}$.

This means that the slightest error (ϵ) in any initial condition (x) can lead to a macroscopic discrepancy (Δ) in the evolution of the dynamics. Accordingly, Δ tells us at what scale these errors show up. Suppose I start a dynamical system in a state x, let it evolve for a while, and try to reproduce this experiment. Even if I reproduce x to within a billionth of an inch, the initial minuscule error may magnify to a large difference in behavior in finite (often relatively short) time, that is, I may find that the second run of the same experiment bears little resemblance to the first. This is what Poincaré meant by his comment quoted in Section 1.1.1.

For linear expanding maps this property is clearly true: Any initial error of an orbit for E_m grows exponentially (by a factor of $|m|$ at every iteration) until it has grown to more than $1/2|m|$. In particular, $\delta = 1/2|m|$ is a sensitivity constant. On the other hand, this property clearly fails for isometries because points do not move apart at all under iteration.

It is important for the definition that Δ does not depend on x, nor on ϵ, but only on the system. Thus, the *smallest* error *anywhere* can lead to discrepancies of size Δ eventually.[2]

Theorem 7.2.12 *Chaotic maps exhibit sensitive dependenceon initial conditions, except when the entire space consists of a single periodic orbit.*

Proof Unless the entire space consists only of a single periodic orbit, the density of periodic points implies that there are two distinct periodic orbits. Since they have no common point, there are periodic points p, q such that $\Delta := \min\{d(f^n(p), f^m(q)) \mid n, m \in \mathbb{Z}\}/8 > 0$. (Note that n and m need not agree.) We now show that Δ is a sensitivity constant.

If $x \in X$, the orbit of one of these two points keeps a distance at least 4Δ from x: If they were both within less than 4Δ of x, then their mutual distance would be less than 8Δ. Suppose this point is q.

Take any $\epsilon \in (0, \Delta)$. By the density of periodic points, there is a periodic point $p \in B(x, \epsilon)$ whose period we call n. Then the set

$$V := \bigcap_{i=0}^{n} f^{-i}(B(f^i(q), \Delta))$$

[2] The meteorologist Edward Lorenz described this as the "butterfly effect": Weather appears to be a chaotic dynamical system, so it is conceivable that a butterfly that flutters by in Rio may cause a typhoon in Tokyo a few days later.

of points whose first n iterates track those of q up to Δ is an open neighborhood of q. By Proposition 7.2.2 (used in the direction that does not require completeness) there exists a $k \in \mathbb{N}$ such that $f^k(B(x,\epsilon)) \cap V \neq \varnothing$, that is, there exists a $y \in B(x,\epsilon)$ such that $f^k(y) \in V$. If $j := \lfloor k/n \rfloor + 1$, then $k/n < j \leq (k/n) + 1$ and

$$k = n \cdot \frac{k}{n} < nj \leq n\left(\frac{k}{n} + 1\right) = k + n.$$

If we take $N := nj$, then this shows that $0 < N - k \leq n$. Since $f^N(p) = p$, the triangle inequality gives

$$\begin{aligned} d(f^N(p), f^N(y)) &= d(p, f^N(y)) \\ &\geq d(x, f^{N-k}(q)) - d(f^{N-k}(q), f^N(y)) - d(p, x) \\ &\geq 4\Delta - \Delta - \Delta = 2\Delta \end{aligned} \tag{7.2.2}$$

because $p \in B(x,\epsilon) \subset B(x,\Delta)$ and

$$f^N(y) = f^{N-k}(f^k(y)) \in f^{N-k}(V) \subset B(f^{N-k}(q), \Delta)$$

by definition of V. Both p and y are in $B(x,\epsilon)$ and either $d(f^N(p), f^N(x)) \geq \Delta$ or $d(f^N(y), f^N(x)) \geq \Delta$ by (7.2.2). □

Remark 7.2.13 There are maps exhibiting sensitive dependence that are not chaotic, such as the linear twist from Section 6.1,1. Here, any point x has arbitrarily nearby points (on a vertical segment through x) that move a considerable distance away after sufficiently many iterates. The set of periodic points consists of those points whose second coordinate is rational and is hence dense. On the other hand, this map is clearly not topologically transitive.

Sensitive dependence can be derived from topological mixing alone, without an assumption on periodic points:

Proposition 7.2.14 *A topologically mixing map (on a space with more than one point) has sensitive dependence.*

Proof Take $\Delta > 0$ such that there are points x_1, x_2 with $d(x_1, x_2) > 4\Delta$. We show that Δ is a sensitivity constant.

Let $V_i = B_\Delta(x_i)$ for $i = 1, 2$. Suppose $x \in X$ and U is a neighborhood of x. By topological mixing there are $N_1, N_2 \in \mathbb{N}$ such that $f^n(U) \cap V_1 \neq \varnothing$ for $n \geq N_1$ and $f^n(U) \cap V_2 \neq \varnothing$ for $n \geq N_2$. For $n \geq N := \max(N_1, N_2)$, there are points $y_1, y_2 \in U$ with $f^n(y_1) \in V_1$ and $f^n(y_2) \in V_2$; hence $d(f^n(y_1), f^n(y_2)) \geq 2\Delta$. By the triangle inequality $d(f^n(y_1), f^n(x)) \geq \Delta$ or $d(f^n(y_2), f^n(x)) \geq \Delta$. □

EXERCISES

Exercise 7.2.1 Find the maximal sensitivity constant for E_2.

Exercise 7.2.2 Find the supremum of sensitivity constants for F_L in Section 7.2.4.

Exercise 7.2.3 Prove that, for a topologically mixing map, any number less than the diameter $\sup\{d(x, y) \mid x, y \in X\}$ of the space X is a sensitivity constant.

Exercise 7.2.4 Consider the linear twist $T\colon S^1 \times [0,1] \to S^1 \times [0,1]$, $T(x,y) = (x + y, y)$ from Section 6.1.1 that was remarked upon in Remark 7.2.13. Prove that it has the following property of *partial topological mixing*: Let $U, V \subset S^1$ be nonempty open sets. Then there exists $N(U,V) \in \mathbb{N}$ such that $T^n(U \times [0,1]) \cap (V \times [0,1]) \neq \varnothing$ for any $n \geq N$.

Exercise 7.2.5 Show that for a compact space sensitive dependence is a topological invariant (see Section 7.3.6).

Exercise 7.2.6 Prove that for any two periodic points of F_L the set of heteroclinic points (see Definition 2.3.4) is dense.

Exercise 7.2.7 Consider a 2×2 integer matrix L without eigenvalues of absolute value 1 and with $|\det L| > 1$. Prove that the induced noninvertible hyperbolic linear map $F_L\colon \mathbb{T}^2 \to \mathbb{T}^2$ is topologically mixing.

7.3 CODING

One of the most important ideas for studying complicated dynamics sounds strange at first. It involves throwing away some information by tracking orbits only approximately. The idea is to divide the phase space into finitely many pieces and to follow an orbit only to the extent of specifying which piece it is in at a given time. This is a bit like the itinerary of the harried tourist in Europe, who decides that it is Tuesday, so the place must be Belgium. A more technological analogy would be to look at the records of a cell phone addict and track which local transmitters were used at various times.

In these analogies one genuinely loses information, because the sequence of European countries or of local cellular stations does not pinpoint the traveller at any given moment. However, orbits in a dynamical system do not move around at whim, and the deterministic nature of the dynamics has the effect that a complete *itinerary* of this sort may (and often does) give all the information about a point. This is the process of *coding* of a dynamical system.

7.3.1 Linear Expanding Maps

The linear expanding maps

$$E_m\colon S^1 \to S^1,\ E_m(x) = mx \pmod 1$$

from Section 7.1.1 are chaotic (Corollary 7.2.8), that is, they exhibit coexistence of dense orbits (Proposition 7.2.7) with a countable dense set of periodic orbits (Proposition 7.1.3). Thus the orbit structure is both complicated and highly nonuniform. Now we look at these maps from a different point of view, which in turn gives a deeper appreciation of just how complicated their orbit structure really is. To simplify notations, assume as before that $m = 2$.

Consider the binary intervals

$$\Delta_n^k := \left[\frac{k}{2^n}, \frac{k+1}{2^n}\right] \qquad \text{for } n = 1, \ldots \text{ and } k = 0, 1, \ldots, 2^n - 1.$$

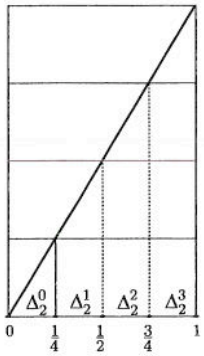


Figure 7.3.1. Linear coding.

Figure 7.3.1 illustrates this for $n = 2$. Let $x = 0.x_1x_2\ldots$ be the binary representation of $x \in [0, 1]$. Then $2x = x_1.x_2x_3\ldots = 0.x_2x_3\ldots \pmod 1$. Thus

$$E_2(x) = 0.x_2x_3\ldots \pmod 1. \tag{7.3.1}$$

This is the first and easiest example of *coding*, which we will discuss in greater detail shortly.

7.3.2 Implications of Coding

We briefly derive a few new facts about linear expanding maps that are best seen via this coding.

1. Proof of Transitivity via Coding. First, we use this representation to give another proof of topological transitivity by describing explicitly the binary representation of a number whose orbit under the iterates of E_2 is dense. Consider an integer k, $0 \le k \le 2^n - 1$. Let $k_0\ldots k_{n-1}$ be the binary representation of k, maybe with several zeroes at the beginning. Then $x \in \Delta_n^k$ if and only if $x_i = k_i$ for $i = 0, \ldots n-1$. Therefore we write $\Delta_{k_0\ldots k_{n-1}} := \Delta_n^k$ from now on. Now put the binary representations of all numbers from 0 to $2^n - 1$ (with zeroes in front if necessary) one after another and form a finite sequence, which we denote by ω_n, that is, ω_n is obtained by concatenating all 2^n binary sequences of length n. Having done this for every $n \in \mathbb{N}$, put the sequences ω_n, $n = 1, 2, \ldots$ in that order, call the resulting infinite sequence ω, and consider the number x with the binary representation $0.\omega$. Since by construction moving ω to the left and cutting off the first digits produces at various moments binary representations of any n-digit number, this means that the orbit of the point x under the iterates of the map E_2 intersects every interval $\Delta_{k_0\ldots k_{n-1}}$ and hence is dense.

This construction extends to any $m \ge 2$. To construct a dense orbit for E_m with $m \le -2$, we notice that $E_m^2 = E_{m^2}$. Obviously the orbit of any point under the iterates of a square of a map is a subset of the orbit under the iterates of the map itself; thus if the former is dense, so is latter. So we apply our construction to the map E_{m^2} and obtain a point with dense orbit under E_m.

2. Exotic Asymptotics. Next we use this approach to show that besides periodic and dense orbits there are other types of asymptotic behavior for orbits of expanding maps. One can construct such orbits for E_2, but the simplest and most elegant example appears for the map E_3.

Proposition 7.3.1 *There exists a point $x \in S^1$ such that the closure of its orbit with respect to the map E_3 in additive notation coincides with the standard middle-third Cantor set K. In particular, K is E_3-invariant and contains a dense orbit.*

Proof The middle-third Cantor set K can be described as the set of all points on the unit interval that have a representation in base 3 with only 0's and 2's as digits (see Section 2.7.1). Similarly to (7.3.1), the map E_3 acts as the shift of digits to the left in the base 3 representation. This implies that K is E_3-invariant. It remains to show that E_3 has a dense orbit in K.

Every point in K has a unique representation in base 3 without 1's. Let $x \in K$ and

$$0.x_1x_2x_3\ldots \tag{7.3.2}$$

be such a representation. Let $h(x)$ be the number whose representation in base 2 is

$$0.\frac{x_1}{2}\frac{x_2}{2}\frac{x_3}{2}\ldots,$$

that is, it is obtained from (7.3.2) by replacing 2's by 1's. Thus we have constructed a map $h\colon K \to [0, 1]$ that is continuous, nondecreasing [that is, $x > y$ implies $h(x) \ge h(y)$], and one-to-one, except for the fact that binary rationals have two preimages each (compare Section 2.7.1 and Section 4.4.1). Furthermore, $h \circ E_3 = E_2 \circ h$. Let $D \subset [0, 1]$ be a dense set of points that does not contain binary rationals. Then $h^{-1}(D)$ is dense in K because, if Δ is an open interval such that $\Delta \cap K \ne \varnothing$, then $h(\Delta)$ is a nonempty interval open, closed, or semiclosed and hence contains points of D. Now take any $x \in [0, 1]$ whose E_2-orbit is dense; the E_3-orbit of $h^{-1}(x) \in K$ is dense in K. □

3. Nonrecurrent Points. Another interesting example is the construction of a nonrecurrent point, that is, such a point x that for some neighborhood U of x all iterates of x avoid U (see Definition 6.1.8). In fact, there is a dense set of nonrecurrent points for the map E_2.

Pick any fixed sequence $(\omega_0, \ldots, \omega_{n-1})$ of 0's and 1's and add a tail of 0's if $\omega_{n-1} = 1$, or of 1's if $\omega_{n-1} = 0$. Call the resulting infinite sequence ω. As before, let x be the number with binary representation $0.\omega$. Thus, x lies in a prescribed interval $\Delta_{\omega_0\ldots\omega_{n-1}}$ and by construction $x \ne 0$. On the other hand, $E_2^n x = 0$ and hence $E_2^m x = 0$ for all $m \ge n$, so x is a nonrecurrent point.

Thus, we have found that E_m is chaotic and topologically mixing, that its periodic and nonrecurrent orbits are dense, and that E_3 has orbits whose closure is a Cantor set.

7.3.3 A Two-Dimensional Cantor Set

We now describe a map in the plane that naturally gives rise to a two-dimensional Cantor set (previously encountered in Problem 2.7.5) on which ternary expansion of the coordinates provides all information about the dynamics. This *horseshoe map* plays a central role in our further development.

Consider a map defined on the unit square $[0,1]\times[0,1]$ by the following construction: First apply the linear transformation $(x,y)\mapsto(3x,y/3)$ to get a horizontal strip whose left and right thirds will be rigid in the next transformation. Holding the left third fixed, bend and stretch the middle third such that the right third falls rigidly on the top third of the original unit square. This results in a "G"-shape. For points that are in and return to the unit square, this map is given analytically by

$$(x,y)\mapsto\begin{cases}(3x,y/3) & \text{if } x\le 1/3\\ (3x-2,(y+2)/3) & \text{if } x\ge 2/3.\end{cases}$$

The inverse can be written as

$$(x,y)\mapsto\begin{cases}(x/3,3y) & \text{if } y\le 1/3\\ ((x+2)/3,3y-2) & \text{if } y\ge 2/3.\end{cases}$$

Geometrically, the inverse looks like an "e"-shape rotated counterclockwise by 90°.

To iterate this map one triples the x-coordinate repeatedly and always assumes that the resulting value is either at most 1/3 or else at least 2/3, that is, that the first ternary digit is 0 or 2, but not 1. (If the expansion is not unique, one requires such a choice to be possible.) Comparing with the construction of the ternary Cantor set in Section 2.7.1, one sees that the x-coordinate lies in the ternary Cantor set C. Looking at the inverse one sees likewise that, in order for all preimages to be defined, the y-coordinate lies in the Cantor set as well. Therefore this map is defined for all positive and negative iterates on the two-dimensional Cantor set $C\times C$. There is a straightforward way of using ternary expansion to code the dynamics. For a point (x,y) the map shifts the ternary expansion of x one step to the left, dropping the first term, and shifts the ternary expansion of y to the right. It is natural to fill in the now-ambiguous first digit of the shifted y-coordinate with the entry from the x-coordinate that was just dropped. This retains all information, and the best way of vizualizing the result is to write the expansion of the y-coordinate in reverse and in front of that of the x-coordinate. This gives a bi-infinite string of 0's and 2's (remember, no 1's allowed), which is shifted by the map. Of course, one should verify that the inverse acts by shifting in the opposite direction.

7.3.4 Sequence Spaces

Now we are ready to discuss the concept of coding in general. We mean by coding a representation of points in the phase space of a discrete-time dynamical system or an invariant subset by sequences (not necessarily unique) of symbols from a certain "alphabet," in this case the symbols $0,\dots,N-1$. So we should acquaint ourselves with these spaces.

Denote by Ω_N^R the space of sequences $\omega=(\omega_i)_{i=0}^{\infty}$ whose entries are integers between 0 and $N-1$. Define a metric by

$$d_\lambda(\omega,\omega'):=\sum_{i=0}^{\infty}\frac{\delta(\omega_i,\omega_i')}{\lambda^i}, \tag{7.3.3}$$

where $\delta(k,l)=1$ if $k\neq l$, $\delta(k,k)=0$, and $\lambda>2$. The same definition can be made for two-sided sequences by summing over $i\in\mathbb{Z}$:

$$d_\lambda(\omega,\omega'):=\sum_{i\in\mathbb{Z}}\frac{\delta(\omega_i,\omega_i')}{\lambda^{|i|}}, \tag{7.3.4}$$

for some $\lambda>3$. This means that two sequences are close if they agree on a long stretch of entries around the origin.

Consider the symmetric *cylinder* defined by

$$C_{\alpha_{1-n}\dots\alpha_{n-1}}:=\{\omega\in\Omega_N\mid \omega_i=\alpha_i \text{ for } |i|<n\}.$$

Fix a sequence $\alpha\in C_{\alpha_{1-n}\dots\alpha_{n-1}}$. If $\omega\in C_{\alpha_{1-n}\dots\alpha_{n-1}}$, then

$$d_\lambda(\alpha,\omega)=\sum_{i\in\mathbb{Z}}\frac{\delta(\alpha_i,\omega_i)}{\lambda^{|i|}}=\sum_{|i|\ge n}\frac{\delta(\alpha_i,\omega_i)}{\lambda^{|i|}}\le\sum_{|i|\ge n}\frac{1}{\lambda^{|i|}}=\frac{1}{\lambda^{n-1}}\frac{2}{\lambda-1}<\frac{1}{\lambda^{n-1}}.$$

Thus $C_{\alpha_{1-n}\dots\alpha_{n-1}}\subset B_{d_\lambda}(\alpha,\lambda^{1-n})$, the λ^{1-n}-ball around α. If $\omega\notin C_{\alpha_{1-n}\dots\alpha_{n-1}}$, then

$$d_\lambda(\alpha,\omega)=\sum_{i\in\mathbb{Z}}\frac{\delta(\alpha_i,\omega_i)}{\lambda^{|i|}}\ge\lambda^{1-n}$$

because $\omega_i\neq\alpha_i$ for some $|i|<n$. Thus $\omega\notin B_{d_\lambda}(\alpha,\lambda^{1-n})$, and the symmetric cylinder is the ball of radius λ^{1-n} around any of its points:

$$C_{\alpha_{1-n}\dots\alpha_{n-1}}=B_{d_\lambda}(\alpha,\lambda^{1-n}). \tag{7.3.5}$$

Therefore, balls in Ω_N are described by specifying a symmetric stretch of entries around the initial one.

For one-sided sequences this discussion works along the same lines [one only needs $\lambda>2$ in (7.3.4)] and λ^{1-n}-balls are described by specifying a string of n initial entries.

Our examples [see (7.3.1)] suggest to represent points in the phase space by sequences in such a way that the sequences representing the image of a point are obtained from those representing the point itself by the shift (translation) of the symbols. In this way the given transformation corresponds to the *shift transformation*

$$\begin{aligned}&\sigma:\Omega_N\to\Omega_N, && (\sigma\omega)_i=\omega_{i+1}\\ &\sigma^R:\Omega_N^R\to\Omega_N^R, && (\sigma^R\omega)_i=\omega_{i+1}.\end{aligned} \tag{7.3.6}$$

We often write σ_N for the shift σ on Ω_N and likewise σ_N^R for σ^R on Ω_N^R. For invertible discrete-time systems, any coding involves sequences of symbols extending in both directions; while for noninvertible systems, one-sided sequences do the job. Section 7.3.7 studies these shifts as dynamical systems.

Among the shift transformations that arise from coding there is also a new kind of combinatorial model for a dynamical system that is described by the possibility or impossibility of certain successions of events.

Definition 7.3.2 Let $A=(a_{ij})_{i,j=0}^{N-1}$ be an $N\times N$ matrix whose entries a_{ij} are either 0's or 1's. (We call such a matrix a 0-1 matrix.) Let

$$\Omega_A:=\{\omega\in\Omega_N\mid a_{\omega_n\omega_{n+1}}=1 \text{ for } n\in\mathbb{Z}\}. \tag{7.3.7}$$

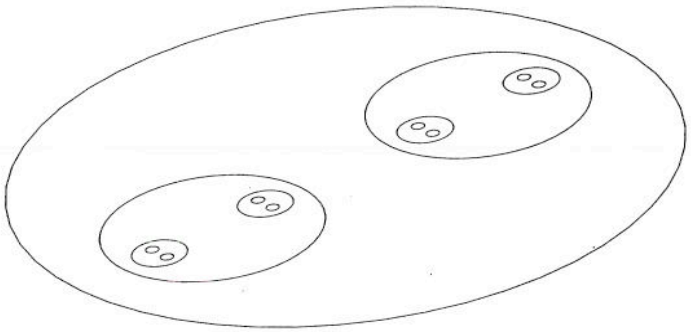


Figure 7.3.2. Obtaining a Cantor set.

The space Ω_A is closed and shift-invariant, and the restriction

$$\sigma_N|_{\Omega_A} =: \sigma_A$$

is called the *topological Markov chain*determined by A.

This is a particular case of a *subshift of finite type.*

7.3.5 Coding

Sequences representing a given point of the phase space are called the *codes* of that point. We have several examples of coding: for the map E_m on the whole circle by sequences from the *alphabet* $\{0, \dots, |m| - 1\}$; for the restriction of the map E_3 to the middle-third Cantor set K by one-sided sequences of 0's and 1's; and for the ternary horseshoe in Section 7.3.3 by bi-infinite sequences of 0's and 2's. In both cases we used one-sided sequences, all sequences appeared as codes of some points, and each code represented only one point. There was, however, an important difference: In the first case, which involved for positive m a representation in base m, a point could have either one or two codes; in the latter there was only one code.

This shows that the space of binary sequences is a Cantor set (Definition 2.7.4). In fact, this also holds for the other sequence spaces.

7.3.6 Conjugacy and Factors

This situation can be roughly described by saying that the shift (Ω_2^R, σ^R) "contains" the map f up to a continuous coordinate change. (We already encountered such a situation in Theorem 4.3.20.)

Definition 7.3.3 Suppose that $g\colon X \to X$ and $f\colon Y \to Y$ are maps of metric spaces X and Y and that there is a continuous surjective map $h\colon X \to Y$ such that $h \circ g = f \circ h$. Then f is said to be a *factor* of g via the *semiconjugacy* or *factor map* h. If this h is a homeomorphism, then f and g are said to be *conjugate* and h is said to be a *conjugacy.*

These notions made a brief appearance in Section 4.3.5 in connection with modeling an arbitrary homeomorphism of the circle by a rotation. The notion of conjugacy is natural and central; two conjugate maps are obtained from one another by a continuous change of coordinates. Hence all properties that are independent of such changes of coordinates are unchanged, such as the numbers of periodic orbits for each period, sensitive dependence (Exercise 7.2.5), topological transitivity, topological mixing, and hence also being chaotic. Such properties are said to be topological invariants. Later in this book we will encounter further important topological invariants such as topological entropy (Definition 8.2.1).

7.3.7 Dynamics of Shifts and Topological Markov Chains

We now study the properties of shifts and topological Markov chains introduced in (7.3.6) and Definition 7.3.2 in more detail. These are important because many interesting dynamical systems are coded by shifts or topological Markov chains. To such dynamical systems the results of this section have immediate applications.

Proposition 7.3.4 *Periodic points for the shifts σ_N and σ_N^R are dense in Ω_N and Ω_N^R, correspondingly, $P_n(\sigma_N) = P_n(\sigma_N^R) = N^n$, and both σ_N and σ_N^R are topologically mixing.*

Proof Periodic orbits for a shift are periodic sequences, that is, $(\sigma_N)^m\, \omega = \omega$ if and only if $\omega_{n+m} = \omega_n$ for all $n \in \mathbb{Z}$. In order to prove density of periodic points, it is enough to find a periodic point in every ball (symmetric cylinder), because every open set contains a ball. To find a periodic point in $C_{\alpha_{-m},\dots,\alpha_m}$, take the sequence ω defined by $\omega_n = \alpha_{n'}$ for $|n'| \le m$, $n' = n \pmod{2m+1}$. It lies in this cylinder and has period $2m+1$.

Every periodic sequence ω of period n is uniquely determined by its coordinates $\omega_0, \dots, \omega_{n-1}$. There are N^n different finite sequences $(\omega_0, \dots, \omega_{n-1})$.

To prove topological mixing, we show that $\sigma_N^n(C_{\alpha_{-m},\dots,\alpha_m}) \cap C_{\beta_{-m},\dots,\beta_m} \neq \varnothing$ for $n > 2m+1$, say, $n = 2m+k+1$ with $k > 0$. Consider any sequence ω such that

$$\omega_i = \alpha_i \text{ for } |i| \le m, \qquad \omega_i = \beta_{i-n} \quad \text{for } i = m+k+1, \dots, 3m+k+1.$$

Then $\omega \in C_{\alpha_{-m},\dots,\alpha_m}$ and $\sigma_N^n(\omega) \in C_{\beta_{-m},\dots,\beta_m}$.

The arguments for the one-sided shift are analogous. □

There is a useful geometric representation of topological Markov chains. Connect i with j by an arrow if $a_{ij} = 1$ to obtain a *Markov graph* G_A with N vertices and several oriented edges. We say that a finite or infinite sequence of vertices of G_A is an *admissible path* or *admissible sequence* if any two consecutive vertices in the sequence are connected by an oriented arrow. A point of Ω_A corresponds to a doubly infinite path in G_A with marked origin; the topological Markov chain σ_A corresponds to moving the origin to the next vertex. The following simple combinatorial lemma is a key to the study of topological Markov chains:

Lemma 7.3.5 *For every $i, j \in \{0, 1, \dots, N-1\}$, the number N_{ij}^m of admissible paths of length $m+1$ that begin at x_i and end at x_j is equal to the entry a_{ij}^m of the matrix A^m.*

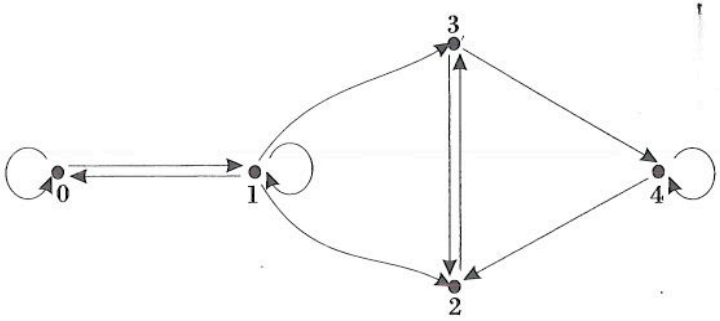


Figure 7.3.3. A Markov graph.

Proof We use induction on m. First, it follows from the definition of the graph G_A that $N_{ij}^1 = a_{ij}$. To show that

$$N_{ij}^{m+1} = \sum_{k=0}^{N-1} N_{ik}^m a_{kj}, \tag{7.3.8}$$

take $k \in \{0, \dots, N-1\}$ and an admissible path of length $m+1$ connecting i and k. It can be extended to an admissible path of length $m+2$ connecting i to j (by adding j) if and only if $a_{kj} = 1$. This proves (7.3.8). Now, assuming by induction that $N_{ij}^m = a_{ij}^m$ for all ij, we obtain $N_{ij}^{m+1} = a_{ij}^{m+1}$ from (7.3.8). □

Corollary 7.3.6 $P_n(\sigma_A) = \operatorname{tr} A^n$.

Proof Every admissible closed path of length $m+1$ with marked origin, that is, a path that begins and ends at the same vertex of G_A, produces exactly one periodic point of σ_A of period m. □

Because the eigenvalue of largest absolute value dominates the trace, it determines the exponential growth rate:

Proposition 7.3.7 $p(\sigma_A) = r(A)$, *where* $r(A)$ *is the spectral radius.*

Proof "$\le$" is clear. To show "$\ge$" we need to avoid cancellations: If $\lambda_j = re^{2\pi i\varphi_j} (1 \le j \le k)$ are the eigenvalues of maximal absolute value then there is a sequence $m_n \to \infty$ such that $m_n\varphi_j \to 0 \pmod 1$ for all j (recurrence for toral translations, Section 5.1), so $\sum \lambda_i^{m_n} \sim r^{m_n}$. □

Example 7.3.8 The Markov graph in Figure 7.3.3 produces three fixed points, $\overline{0}$, $\overline{1}$, and $\overline{4}$. $\overline{01}$ and $\overline{23}$ give four periodic points with period 2. The period-3 orbits are generated by $\overline{011}$, $\overline{001}$, $\overline{234}$.

Topological Markov chains can be classified according to the recurrence properties of various orbits they contain. Now we concentrate on those topological Markov chains that possess the strongest recurrence properties.

Definition 7.3.9 A matrix A is said to be *positive* if all its entries are positive. A 0-1 matrix A is said to be *transitive* if A^m is positive for some $m \in \mathbb{N}$. A topological Markov chain σ_A is said to be *transitive* if A is a transitive matrix.

Lemma 7.3.10 *If* A^m *is positive, then so is* A^n *for any* $n \ge m$.

Proof If $a_{ij}^n > 0$ for all i, j, then for each j there is a k such that $a_{kj} = 1$. Otherwise, $a_{ij}^n = 0$ for every n and i. Now use induction. If $a_{ij}^n > 0$ for all i, j, then $a_{ij}^{n+1} = \sum_{k=0}^{N-1} a_{ik}^n a_{kj} > 0$ because $a_{kj} = 1$ for at least one k. □

Lemma 7.3.11 *If* A *is transitive and* $\alpha_{-k}, \dots, \alpha_k$ *is admissible, that is,* $a_{\alpha_i\alpha_{i+1}} = 1$ *for* $i = -k, \dots, k-1$, *then the intersection* $\Omega_A \cap C_{\alpha_{-k},\dots,\alpha_k} =: C_{\alpha_{-k},\dots,\alpha_k,A}$ *is nonempty and moreover contains a periodic point.*

Proof Take m such that $a_{\alpha_k,\alpha_{-k}}^m > 0$. Then one can extend the sequence α to an admissible sequence of length $2k+m+1$ that begins and ends with α_{-k}. Repeating this sequence periodically, we obtain a periodic point in $C_{\alpha_{-k},\dots,\alpha_k,A}$. □

Proposition 7.3.12 *If* A *is a transitive matrix, then the topological Markov chain* σ_A *is topologically mixing and its periodic orbits are dense in* Ω_A; *in particular,* σ_A *is chaotic and hence has sensitive dependence on initial conditions.*

Proof The density of periodic orbits follows from Lemma 7.3.11. To prove topological mixing, pick open sets $U, V \subset \Omega_A$ and nonempty symmetric cylinders $C_{\alpha_{-k},\dots,\alpha_k,A} \subset U$ and $C_{\beta_{-k},\dots,\beta_k,A} \subset V$. Then it suffices to show that $\sigma_A^n(C_{\alpha_{-k},\dots,\alpha_k,A}) \cap C_{\beta_{-k},\dots,\beta_k,A} \ne \varnothing$ for any sufficiently large n. Take $n = 2k+1+m+l$ with $l \ge 0$, where m is as in Definition 7.3.9. Then $a_{\alpha_k\beta_{-k}}^{m+l} > 0$ by Lemma 7.3.10, so there is an admissible sequence of length $4k+2+m+l$ whose first $2k+1$ symbols are identical to $\alpha_{-k}, \dots, \alpha_k$ and the last $2k+1$ symbols to $\beta_{-k}, \dots, \beta_k\beta$. By Lemma 7.3.11, this sequence can be extended to a periodic element of Ω_A which belongs to $\sigma_A^n(C_{\alpha_{-k},\dots,\alpha_k,A}) \cap C_{\beta_{-k},\dots,\beta_k,A}$. □

Example 7.3.13 The matrix $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ is not transitive because all its powers are upper triangular and hence there is no path from 1 to 0. In fact, the space Ω_A is countable and consists of two fixed points $(\dots, 0, \dots, 0, \dots)$ and $(\dots, 1, \dots, 1, \dots)$, and a single heteroclinic orbit connecting them (consisting of the sequences that are 1 up to some place and 0 thereafter).

Example 7.3.14 For the matrix

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

every orbit alternates between entries from the first group $\{0, 1\}$ on the one hand and from the second group $\{2, 3\}$ on the other hand, that is, the parity (even–odd) must alternate. Therefore no power of the matrix has all entries positive.

EXERCISES

Exercise 7.3.1 Prove that E_2 has a nonperiodic orbit all of whose even iterates lie in the left half of the unit interval.

Exercise 7.3.2 Prove that E_2 has a uncountably many orbits for which no segment of length 10 has more than one point in the left half of the unit interval.

Exercise 7.3.3 Prove that linear maps that are conjugate in the sense of linear algebra are topologically conjugate in the sense of Definition 7.3.3.

Exercise 7.3.4 Write down the Markov matrix for Figure 7.3.3 and check Corollary 7.3.6 up to period 3.

Exercise 7.3.5 Consider the metric

$$d'_\lambda(\alpha, \omega) := \sum_{i\in\mathbb{Z}} \frac{|\alpha_i - \omega_i|}{\lambda^{|i|}} \tag{7.3.9}$$

on Ω_N. Show that for $\lambda > 2N - 1$ the cylinder $C_{\alpha_{1-n}\ldots\alpha_{n-1}}$ is a λ^{1-n}-ball for d'_λ.

Exercise 7.3.6 Repeat the previous exercise for one-sided shifts (with $\lambda > N$).

Exercise 7.3.7 Consider the metric

$$d''_\lambda(\alpha, \omega) := \lambda^{-\max\{n\in\mathbb{N}\,|\,\alpha_i=\omega_i \text{ for } |i|\le n\}} \tag{7.3.10}$$

[and $d''_\lambda(\alpha, \alpha) = 0$] on Ω_N. Show that the cylinder $C_{\alpha_{1-n}\ldots\alpha_{n-1}}$ is a ball for d''_λ.

Exercise 7.3.8 Find the supremum of sensitivity constants for a transitive topological Markov chain with respect to the metric d''_λ.

Exercise 7.3.9 Find the supremum of sensitivity constants for a transitive topological Markov chain with respect to the metric d'_λ.

Exercise 7.3.10 Show that for $m < n$ the shift on Ω_m is a factor of the shift on Ω_n.

Exercise 7.3.11 Prove that the quadratic map f_4 on $[0, 1]$ is not conjugate to any of the maps f_λ for $\lambda \in [0, 4)$.

Exercise 7.3.12 Show that the topological Markov chains determined by the matrices

$$\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

are conjugate.

Exercise 7.3.13 Find the smallest positive value of $p(\sigma_A)$ for a transitive topological Markov chain with two states (that is, with a 2×2 matrix A).

PROBLEMS FOR FURTHER STUDY

Problem 7.3.14 Find all factors of an irrational rotation R_α of the circle.

Problem 7.3.15 Find the smallest value of $p(\sigma_A)$ for a transitive topological Markov chain with three states (that is, with a 3×3 matrix A).

7.4 MORE EXAMPLES OF CODING

We now carry out a coding construction for several familiar dynamical systems.

7.4.1 Nonlinear Expanding Maps

There is a correspondence between general (not necessarily linear) expanding maps of the circle (Section 7.1.3) and a shift on a sequence space. The construction is similar to the one from Section 7.3.1. There is some effort involved, but there is a beautiful prize at the end: We obtain a complete classification of a large class of maps in terms of a simple invariant.

To keep notations simple, we consider an expanding map $f\colon S^1 \to S^1$ of degree 2. By Proposition 7.1.9, f has exactly one fixed point p. (For maps of higher degree, we could pick any one of the fixed points.) Since $\deg(f) = 2$, there is exactly one point $q \neq p$ such that $f(q) = p$. The points p and q divide the circle into two arcs. Starting from p in the positive direction, denote the first arc by Δ_0 and the second arc by Δ_1. Define the coding for $x \in S^1$ as follows: x is represented by the sequence $\omega \in \Omega_2^R$ for which

$$f^n(x) \in \Delta_{\omega_n}. \tag{7.4.1}$$

This representation is unique unless $f^n(x) \in \{p, q\} = \Delta_0 \cap \Delta_1$. This lack of uniqueness is similar to the case of binary rationals for the map E_2. Suppose a point x has an iterate in $\{p, q\}$. Then either $x = p$ and $f^n(x) = p$ for all $n \in \mathbb{N}$, or else the point q must appear before p in the sequence of iterates, that is, $f^n(x) \notin \{p, q\}$ for all n less than some k and then $f^k(x) = q$ and $f^{k+1}(x) = p$. In this case we make the following convention. p has two codes, all 0's and all 1's, and q has two codes, 01111111... and 1000000..., and any x such that $F^k(x) = q$ has two codes given by the first $k - 1$ digits uniquely defined by (7.4.1), followed by either of the codes for q.

Actually, going the other way around is better:

Proposition 7.4.1 *If $f\colon S^1 \to S^1$ is an expanding map of degree 2, then f is a factor of σ^R on Ω_2^R (Definition 7.3.3), that is, there is a surjective continuous map $h\colon \Omega_2^R \to S^1$ such that $f^n(h(\omega)) \in \Delta_{\omega_n}$ for all $n \in \mathbb{N}_0$, that is, $h \circ \sigma^R = f \circ h$.*

Proof That the domain of h is Ω_2^R requires that every sequence of 0's and 1's appears as the code of some point. First, f maps each of the two intervals Δ_0 and Δ_1 onto S^1 almost injectively, the only identification being at the ends. Let

Δ_{00} be the core of $\Delta_0 \cap f^{-1}(\Delta_0)$,

Δ_{01} be the core of $\Delta_0 \cap f^{-1}(\Delta_1)$,

Δ_{10} be the core of $\Delta_1 \cap f^{-1}(\Delta_0)$,

Δ_{11} be the core of $\Delta_1 \cap f^{-1}(\Delta_1)$.

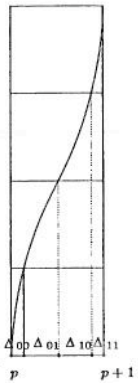


Figure 7.4.1. Nonlinear coding.

What we mean by "core" is that each indicated intersection consists of an interval as well as an isolated point (p or q), and we discard this extraneous point. Each of these four intervals is mapped onto S^1 by f^2, again the only identification being at the ends. By definition, any point from Δ_{ij} has ij as the first two symbols of its code. Proceeding inductively we construct for any finite sequence $\omega_0, \dots, \omega_{n-1}$ the interval

$$\Delta_{\omega_0,\dots,\omega_{n-1}} := \text{the core of } \Delta_{\omega_0} \cap f^{-1}(\Delta_{\omega_1}) \cdots \cap f^{1-n}(\Delta_{\omega_{n-1}}), \tag{7.4.2}$$

which is mapped by f^n onto S^1 with identification of the endpoints. Now take any infinite sequence $\omega = \omega_1, \dots \in \Omega_2^R$. The intersection $\bigcap_{n=1}^{\infty} \Delta_{\omega_0,\dots,\omega_{n-1}}$ of the nested closed intervals $\Delta_{\omega_0,\dots,\omega_{n-1}}$ is nonempty, and any point in this intersection has the sequence ω as its code.

So far we have only used the fact that f is a monotone map of degree 2. To show that h is well defined, we use the expanding property to check that $\bigcap_{n=1}^{\infty} \Delta_{\omega_0,\dots,\omega_{n-1}}$ consists of a single point, hence a point with a given code is unique.

If $g\colon I \to S^1$ is an injective map of an open interval I with a nonnegative derivative, then by the Mean-Value Theorem A.2.3 $l(g(I)) = \int_I g'(x)\,dx = g'(\xi) l(I)$ for some $\xi \in I$. Thus, in our case, there is a ξ_n such that

$$1 = l(S^1) = \int_{\Delta_{\omega_0,\dots,\omega_{n-1}}} (f^n)'(x)\,dx = (f^n)'(\xi_n) \cdot l(\Delta_{\omega_0,\dots,\omega_{n-1}}).$$

Since f is expanding $|(f^n)'| > \lambda^n$ for some $\lambda > 1$, hence $l(\Delta_{\omega_0,\dots,\omega_{n-1}}) < \lambda^{-n} \to 0$ as $n \to \infty$ and $\bigcap_{n=1}^{\infty} \Delta_{\omega_0,\dots,\omega_{n-1}}$ consists of a single point x_ω. This gives a well-defined surjective map $h\colon \Omega_2^R \to S^1$, $\omega \mapsto x_\omega$.

Give Ω_2^R the metric d_4 from (7.3.3). We showed in Section 7.3.4 that if $\epsilon = \lambda^{-n}$ and $\delta = 4^{-n}$, then $d(\omega, \omega') < \delta$ implies that $\omega_i = \omega_i'$ for $i < n$ and hence $|x_\omega - x_{\omega'}| \le l(\Delta_{\omega_0,\dots,\omega_{n-1}}) < \lambda^{-n} = \epsilon$. Thus h is continuous.

That $h(\sigma^R(\omega)) = f(h(\omega))$ is clear from the construction. □

7.4.2 Classification via Coding

Proposition 7.4.1 and the discussion preceding it established a semiconjugacy between the one-sided 2-shift and the expanding map f on S^1, that is,

Proposition 7.4.2 *Let $f\colon S^1 \to S^1$ be an expanding map of degree 2. Then f is a factor of the one-sided 2-shift (Ω_2^R, σ_R) via a semiconjugacy $h\colon \Omega_2^R \to S^1$. If $h(\omega) = h(\omega') =: x$, then there exists an $n \in \mathbb{N}_0$ such that $f^n(x) \in \{p, q\}$, where $p = f(p) = f(q)$, $q \neq p$.*

The last sentence of this proposition says that h is "very close" to being a conjugacy: There are only countably many image points where injectivity fails.

An important feature of this coding is that it is obtained in a uniform way for all expanding maps, and that the absence of injectivity occurs at points defined by their dynamics, namely, the fixed point and its preimages. This leads us to the prize promised at the beginning:

Theorem 7.4.3 *If $f, g\colon S^1 \to S^1$ are expanding maps of degree 2, then f and g are topologically conjugate; in particular, every expanding map of S^1 of degree 2 is conjugate to E_2.*

Proof We have semiconjugacies $h_f, h_g\colon \Omega_2^R \to S^1$ for f and g. For $x \in S^1$, consider the set $H_x := h_g(h_f^{-1}(\{x\}))$. If x is a point of injectivity of h_f, that is, $h_f^{-1}(\{x\})$ is a single point, then so is H_x. Otherwise, x is a preimage of the fixed point under some iterate of f and $h_f^{-1}(\{x\})$ consists of a collection of sequences that are mapped under h_g to a single point. Therefore, H_x always consists of precisely one point $h(x)$. The bijective map $h\colon S^1 \to S^1$ thus defined is clearly a conjugacy: $h \circ f = g \circ h$. It is continuous because h_f sends open sets to open sets, that is, the image of a sequence and all sufficiently closeby sequences contains a small interval. Exchanging f and g shows that h^{-1} is also continuous. □

This holds for any degree via an appropriate coding. It is the first major conjugacy result that establishes conjugacy with a specific model for all maps from a certain class. The Poincaré Classification Theorem 4.3.20 comes close, but requires extra assumptions (such as the existence of the second derivative; see Section 4.4.3) to produce a conjugacy with a rotation.

7.4.3 Quadratic Maps

For $\lambda > 4$ consider the quadratic map

$$f\colon \mathbb{R} \to \mathbb{R}, \quad x \to \lambda x(1-x).$$

If $x < 0$, then $f(x) < x$ and $f'(x) > \lambda > 4$, so $f^n(x) \to -\infty$. When $x > 1$, $f(x) < 0$ and hence $f^n(x) \to -\infty$. Thus the set of points with bounded orbits is $\bigcap_{n\in\mathbb{N}_0} f^{-n}([0,1])$.

Proposition 7.4.4 *If $\lambda > 2 + \sqrt{5}$ and $f\colon \mathbb{R} \to \mathbb{R}$, $x \to \lambda x(1-x)$, then there is a homeomorphism $h\colon \Omega_2^R \to \Lambda := \bigcap_{n\in\mathbb{N}_0} f^{-n}([0,1])$ such that $h \circ \sigma^R = f \circ h$, that is, $f_{|\Lambda}$ is conjugate to the 2-shift.*

Proof Let

$$\Delta_0 = \left[0, \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{1}{\lambda}}\right] \quad \text{and} \quad \Delta_1 = \left[\frac{1}{2} + \sqrt{\frac{1}{4} - \frac{1}{\lambda}}, 1\right].$$

Then $f^{-1}([0,1]) = \Delta_0 \cup \Delta_1$ by solving the quadratic equation $f(x) = 1$. Likewise, $f^{-2}([0,1]) = \Delta_{00} \cup \Delta_{01} \cup \Delta_{11} \cup \Delta_{10}$ consists of four intervals, and so forth. Consider the partition of Λ by Δ_0 and Δ_1. These pieces do not overlap and

$$|f'(x)| = |\lambda(1-2x)| = 2\lambda\left|x - \frac{1}{2}\right| \geq 2\lambda\sqrt{\frac{1}{4} - \frac{1}{\lambda}}$$

$$= \sqrt{\lambda^2 - 4\lambda} > \sqrt{(2+\sqrt{5})^2 - 4(2+\sqrt{5})} = 1$$

on $\Delta_0 \cup \Delta_1$. Thus, for any sequence $\omega = (\omega_0, \omega_1, \dots)$, the diameter of the intersections

$$\bigcap_{n=0}^{N} f^{-n}(\Delta_{\omega_n})$$

decreases (exponentially) as $N \to \infty$. This shows that for a sequence $\omega = (\omega_0, \omega_1, \dots)$ the intersection

$$h(\{\omega\}) = \bigcap_{n \in \mathbb{N}_0} f^{-n}(\Delta_{\omega_n}) \tag{7.4.3}$$

consists of exactly one point and this map $h\colon \Omega_2^R \to \Lambda$ is a homeomorphism. □

Remark 7.4.5 It turns out that Proposition 7.4.4 holds whenever $\lambda > 4$ (Proposition 11.4.1), but this is significantly less straightforward to prove than the present result. The situation present in either case, where a map folds an interval entirely over itself, is referred to as a one-dimensional *horseshoe*, in analogy to the geometry seen in the next subsection.

7.4.4 Linear Horseshoe

We now describe Smale's original "*horseshoe*," which provides one of the best examples of perfect coding. (In Section 7.3.3 a special case was constructed, in which ternary expansion provides the coding.)

Let Δ be a rectangle in $\mathbb{R}^2$ and $f\colon \Delta \to \mathbb{R}^2$ a diffeomorphism of Δ onto its image such that the intersection $\Delta \cap f(\Delta)$ consists of two "horizontal" rectangles Δ_0 and Δ_1 and the restriction of f to the components $\Delta^i := f^{-1}(\Delta_i)$, $i = 0, 1$, of $f^{-1}(\Delta)$ is a hyperbolic linear map, contracting in the vertical direction and expanding in the horizontal direction. This implies that the sets Δ^0 and Δ^1 are "vertical" rectangles. One of the simplest ways to achieve this effect is to bend Δ into a "horseshoe", or rather into the shape of a permanent magnet (Figure 7.4.2), although this method produces some inconveniences with orientation. Another way, which is better from the point of view of orientation, is to bend Δ roughly into a paper clip shape (Figure 7.4.3). This is an exaggerated version of the ternary horseshoe in Section 7.3.3, which also leaves some extra margin. If the horizontal and vertical rectangles lie strictly inside Δ, then the maximal invariant subset $\Lambda = \bigcap_{n=-\infty}^{\infty} f^{-n}(\Delta)$ of Δ is contained in the interior of Δ.

Proposition 7.4.6 $f_{\restriction \Lambda}$ *is topologically conjugate to* σ_2.

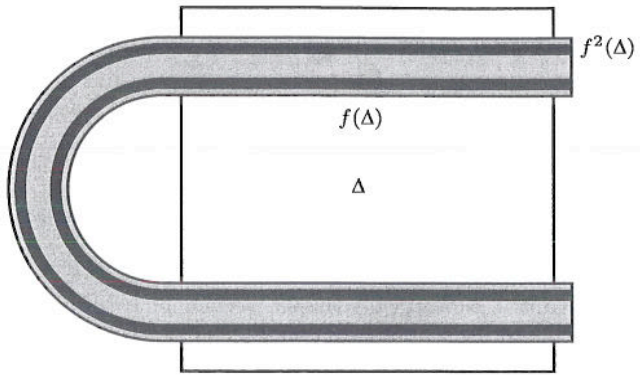


Figure 7.4.2. The horseshoe.

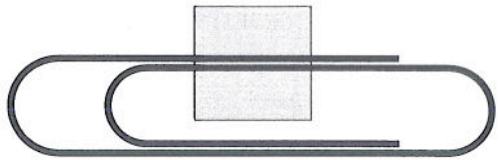


Figure 7.4.3. The paper clip.

Proof We use Δ^0 and Δ^1 as the "pieces" in the coding construction and start with positive iterates. The intersection $\Delta \cap f(\Delta) \cap f^2(\Delta)$ consists of four thin horizontal rectangles: $\Delta_{ij} = \Delta_i \cap f(\Delta_j) = f(\Delta^i) \cap f^2(\Delta^j)$, $i, j \in \{0, 1\}$ (see Figure 7.4.2). Continuing inductively, one sees that $\bigcap_{i=0}^{n} f^i(\Delta)$ consists of 2^n thin disjoint horizontal rectangles whose heights are exponentially decreasing with n. Each such rectangle has the form $\Delta_{\omega_1,\dots,\omega_n} = \bigcap_{i=1}^{n} f^i(\Delta^{\omega_i})$, where $\omega_i \in \{0, 1\}$ for $i = 1, \dots, n$. Each infinite intersection $\bigcap_{n=1}^{\infty} f^n(\Delta^{\omega_n})$, $\omega_n \in \{0, 1\}$, is a horizontal segment, and the intersection $\bigcap_{n=1}^{\infty} f^n(\Delta)$ is the product of the horizontal segment with a Cantor set in the vertical direction. Similarly, one defines and studies vertical rectangles $\Delta^{\omega_0,\dots,\omega_{-n}} = \bigcap_{i=0}^{n} f^{-i}(\Delta^{\omega_{-i}})$, the vertical segments $\bigcap_{n=0}^{\infty} f^{-n}(\Delta^{\omega_{-n}})$, and the set $\bigcap_{n=0}^{\infty} f^{-n}(\Delta)$, which is the product of a segment in the vertical direction with a Cantor set in the horizontal direction.

The desired invariant set $\Lambda = \bigcap_{n=-\infty}^{\infty} f^{-n}(\Delta)$ is the product of two Cantor sets and hence is a Cantor set itself (Problem 2.7.5), and the map

$$h\colon \Omega_2 \to \Lambda, \qquad h(\{\omega\}) = \bigcap_{n=-\infty}^{\infty} f^{-n}(\Delta^{\omega_n})$$

is a homeomorphism that conjugates the shift σ_2 and the restriction of the diffeomorphism f to the set Λ. □

Since periodic points and topological mixing are invariants of topological conjugacy, Proposition 7.4.6 and Proposition 7.3.4 immediately give substantial information about the behavior of f on Λ.

Corollary 7.4.7 *Periodic points of f are dense in Λ, $P_n(f|_\Lambda) = 2^n$, and the restriction of f to Λ is topologically mixing.*

Remark 7.4.8 Any map for which there is a perfect coding is defined on a Cantor set, because the perfect coding establishes a homeomorphism between the phase space and a sequence space, which is a Cantor set.

7.4.5 Coding of the Toral Automorphism

The idea of coding can be applied to hyperbolic toral automorphisms. To simplify notations and keep the construction more visual, we consider the standard example. Among our examples, this is the first where the coding is ingenious, even though it is geometrically simple. Section 10.3 describes a construction whose dynamical implications are quite similar to those obtained here, but where the geometry is complicated and almost always fractal.

Theorem 7.4.9 *For the map*

$$F(x, y) = (2x + y,\ x + y) \qquad (\text{mod } 1)$$

of the 2-torus from Section 7.1.4 there is a semiconjugacy $h\colon \Omega_A \to \mathbb{T}^2$ with

$$F \circ h = h \circ \sigma_5|_{\Omega_A}, \qquad \textit{where}$$

$$A = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}. \tag{7.4.4}$$

Proof Draw segments of the two eigenlines at the origin until they cross sufficiently many times and separate the torus into disjoint rectangles. Specifically, extend a segment of the contracting line in the fourth quadrant until it intersects the segment of the expanding line twice in the first quadrant and once in the third quadrant (see Figure 7.4.4). The resulting configuration is a decomposition of the torus into two rectangles $R^{(1)}$ and $R^{(2)}$. Three pairs among the seven vertices of the plane configuration are identified, so there are only four different points on the torus that serve as vertices of the rectangles; the origin and three intersection points. Although $R^{(1)}$ and $R^{(2)}$ are not disjoint, one can apply the method used for the horseshoe, using $R^{(1)}$ and $R^{(2)}$ as basic rectangles. The expanding and contracting eigendirections play the role of the "horizontal" and "vertical" directions, correspondingly. Figure 7.4.5 shows that the image $F(R^{(i)})$ $(i = 1, 2)$ consists of several "horizontal" rectangles of full length. The union of the boundaries $\partial R^{(1)} \cup \partial R^{(2)}$ consists of the segments of the two eigenlines at the origin just described. The image of the contracting segment is a part of that segment. Thus, the images of $R^{(1)}$ and $R^{(2)}$ have to be "anchored" at parts of their "vertical" sides; that is, once one of the images "enters" either $R^{(1)}$ or $R^{(2)}$, it has to stretch all the way through it. By matching things up along the contracting direction one sees that $F(R^{(1)})$ consists of three components, two in $R^{(1)}$ and one in $R^{(2)}$. The image of $R^{(2)}$ has two components, one in each

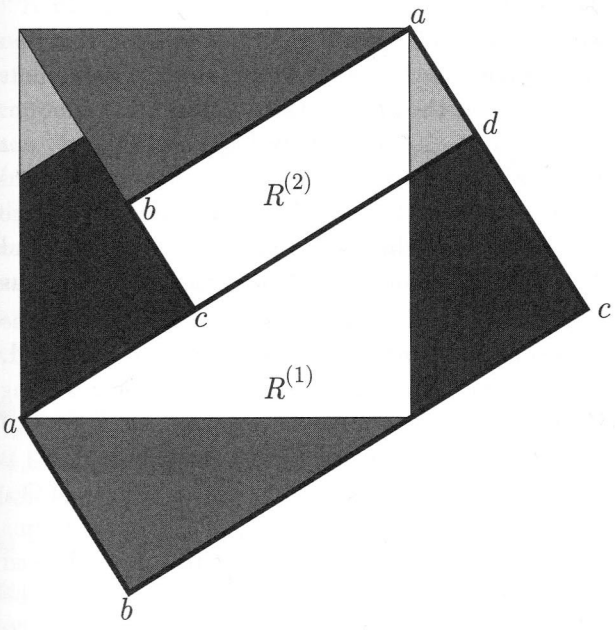


Figure 7.4.4. Partitioning the torus.

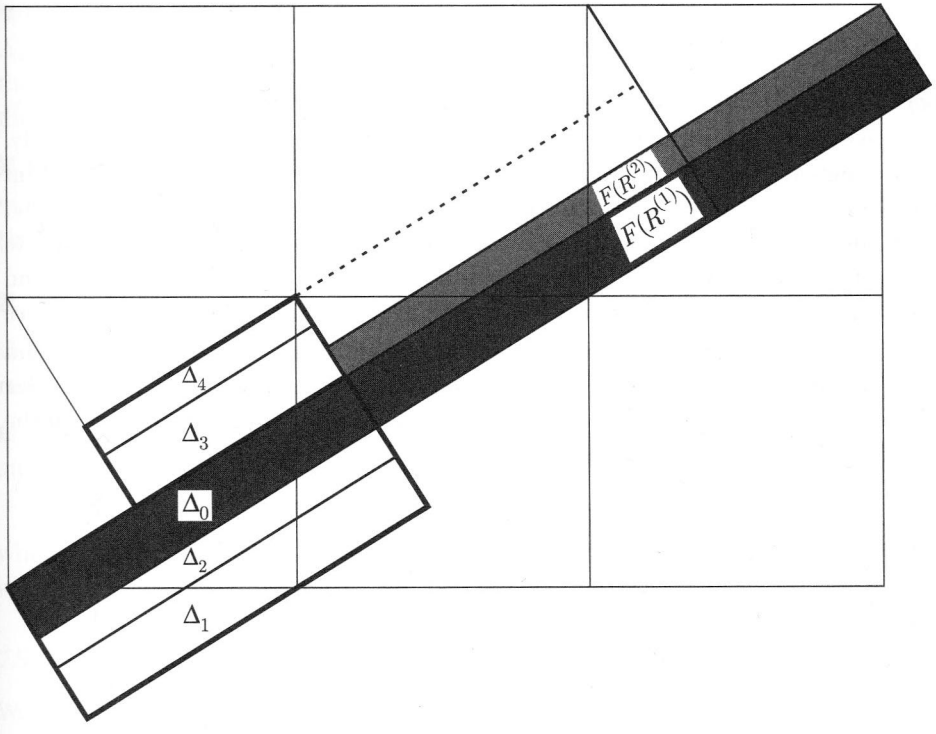


Figure 7.4.5. The image of the partition.

rectangle (see Figure 7.4.5). The fact that $F(R^{(1)})$ has two components in $R^{(1)}$ would cause problems if we were to use $R^{(1)}$ and $R^{(2)}$ for coding construction (more than one point for some sequences), but we use these five components Δ_0, Δ_1, Δ_2, Δ_3, Δ_4 (or their preimages) as the pieces in our coding construction. There is exactly one rectangle $\Delta_{\omega_{-\ell}\dots\omega_0,\omega_1\dots\omega_k}$ defined by $\bigcap_{n=-\ell}^{k} F^{-n}(\Delta_{\omega_n})$, not several. (As in the case of expanding maps in Section 7.3.1, we have to discard extraneous pieces, in this case line segments.) Due to the contraction of F in the "vertical" direction, $\Delta_{\omega_{-\ell}\dots\omega_0,\omega_1\dots\omega_k}$ has "height" less than $((3-\sqrt{5})/2)^{\ell}$, and due to the contraction of F^{-1} in the "horizontal" direction $\Delta_{\omega_{-\ell}\dots\omega_0,\omega_1\dots\omega_k}$ has "width" less than $((3-\sqrt{5})/2)^{k}$. These go to zero as $\ell \to \infty$ and $k \to \infty$, so the intersection $\bigcap_{n\in\mathbb{Z}} F^{-n}(\Delta_{\omega_n})$ defines at most one point $h(\omega)$. On the other hand, because of the "Markov" property described previously, that is, the images going full length through rectangles, the following is true: If $\omega \in \Omega_5$ and $F^{-1}(\Delta_{\omega_n})$ overlaps $\Delta_{\omega_{n+1}}$ for all $n \in \mathbb{Z}$, then there is such a point $h(\omega)$ in $\bigcap_{n\in\mathbb{Z}} F^{-n}(\Delta_{\omega_n})$. Thus, we have a coding, which, however, is not defined for all sequences of Ω_5.

Instead, we have to restrict attention to the subspace Ω_A of Ω_5 that contains only those sequences where any two successive entries constitute an "allowed transition", that is, 0, 1, 2 can be followed by 0, 1, or 3, and 3 and 4 can be followed by 2 or 4. This is exactly the topological Markov chain (Definition 7.3.2) for (7.4.4). □

Theorem 7.4.10 *The semiconjugacy between σ_A and F is one-to-one on all periodic points except for the fixed points. The number of preimages of any point not negatively asymptotic to the fixed point is bounded.*

Proof We describe carefully the identifications arising from our semiconjugacy, that is, what points on the torus have more than one preimage. First, obviously, the topological Markov chain σ_A has three fixed points, namely, the constant sequences of 0's, 1's, and 4's, whereas the toral automorphism F has only one, the origin. It is easy to see that all three fixed points are indeed mapped to the origin. As we have seen in Proposition 7.1.10, $P_n(F) = \lambda_1^n + \lambda_1^{-n} - 2$, and accordingly $P_n(\sigma_A) = \operatorname{tr} A^n = \lambda_1^n + \lambda_1^{-n} = P_n(F) + 2$ (Corollary 7.3.6), where $\lambda_1 = (3+\sqrt{5})/2$ is the maximal eigenvalue for both the 2×2 matrix $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ and for the 5×5 matrix (7.4.4). To see that the eigenvalues are the same, consider $A - \lambda\,\mathrm{Id}$, subtract column 4 from the first two columns and column 5 from the third, and then add rows 1 and 2 to row 4 and row 3 to row 5:

$$\begin{pmatrix} 1-\lambda & 1 & 0 & 1 & 0 \\ 1 & 1-\lambda & 0 & 1 & 0 \\ 1 & 1 & -\lambda & 1 & 0 \\ 0 & 0 & 1 & -\lambda & 1 \\ 0 & 0 & 1 & 0 & 1-\lambda \end{pmatrix} \to \begin{pmatrix} -\lambda & 0 & 0 & 1 & 0 \\ 0 & -\lambda & 0 & 1 & 0 \\ 0 & 0 & -\lambda & 1 & 0 \\ \lambda & \lambda & 0 & -\lambda & 1 \\ 0 & 0 & \lambda & 0 & 1-\lambda \end{pmatrix}$$

$$\to \begin{pmatrix} -\lambda & 0 & 0 & 1 & 0 \\ 0 & -\lambda & 0 & 1 & 0 \\ 0 & 0 & -\lambda & 1 & 0 \\ 0 & 0 & 0 & 2-\lambda & 1 \\ 0 & 0 & 0 & 1 & 1-\lambda \end{pmatrix}.$$

Furthermore, one can see that every point $q \in \mathbb{T}^2$ whose positive and negative iterates avoid the boundaries $\partial R^{(1)}$ and $\partial R^{(2)}$ has a unique preimage, and vice versa. In particular, periodic points other than the origin (which have rational coordinates) fall into this category. The points of Ω_A whose images are on those boundaries or their iterates under F fall into three categories corresponding to the three segments of stable and unstable manifolds through 0 that define parts of the boundary. Thus sequences are identified in the following cases: They have a constant infinite right (future) tail consisting of 0's or 4's, and agree otherwise – this corresponds to a stable boundary piece – or else an infinite left (past) tail (of 0's and 1's, or of 4's), and agree otherwise – this corresponds to an unstable boundary piece. □

■ EXERCISES

■ **Exercise 7.4.1** Prove that for $\lambda \geq 1$ every bounded orbit of the quadratic map f_λ is in $[0, 1]$.

■ **Exercise 7.4.2** Give a detailed argument that (7.4.3) defines a homeomorphism.

■ **Exercise 7.4.3** Construct a Markov partition for $\left(\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ that consists of two squares.

■ **Exercise 7.4.4** Construct a Markov partition and describe the corresponding topological Markov chain for the automorphism F_L, where $L = \left(\begin{smallmatrix} 1 & 1 \\ 2 & 1 \end{smallmatrix}\right)$.

■ **Exercise 7.4.5** Given a 0-1 $n \times n$-matrix A, describe a system of n rectangles $\Delta_1, \ldots, \Delta_n$ in $\mathbb{R}^2$ and map $f\colon \Delta := \bigcup_{i=1}^n \Delta_i \to \mathbb{R}^2$ such that the restriction of f to the set of points that stay inside Δ for all iterates of f is topologically equivalent to the topological Markov chain σ_A.

■ **Exercise 7.4.6** Check that the process (7.4.2) of discarding extraneous points in the coding construction amounts to taking $\Delta_{\omega_0,\ldots,\omega_{n-1}} = \overline{\bigcap_{i=0}^{n-1} \mathrm{Int}(f^{-i}(\Delta_{\omega_i}))}$, and $\{h(\omega)\} := \bigcap_{n\in\mathbb{N}} \Delta_{\omega_0,\ldots,\omega_{n-1}}$.

■ PROBLEMS FOR FURTHER STUDY

■ **Problem 7.4.7** Show that the assertion of Theorem 7.4.3 remains true for any map f of degree 2 such that $f' \geq 1$ and $f' = 1$ only at finitely many points.

■ **Problem 7.4.8** Prove the assertion of Theorem 7.4.9 for some 0-1 matrix A for any automorphism

$$F_L\colon \mathbb{T}^2 \to \mathbb{T}^2, x \mapsto Lx \pmod 1,$$

where L is an integer 2×2 matrix with determinant $+1$ or -1 and with real eigenvalues different from ± 1.

7.5 UNIFORM DISTRIBUTION

We now investigate whether the notion of the uniform distribution of orbits that appeared in previous chapters for rotations of the circle and translations of the torus has any meaning for the group of examples discussed in the present chapter, such as linear or nonlinear expanding maps of the circle, shifts, and automorphisms of the torus.

the form $\bar{z}^{q-1}$ occur *in addition* to those present when β is irrational. If $q \geq 5$ (weak resonance) it can be shown that these additional resonances play a similar, but not determining, role for all q. However, this is no longer true for $q \leq 4$ (strong resonance) where the additional resonances lead to bifurcations that are characterised by the value of q.

Following Arnold & Takens (see Arnold, 1983, pp. 292—313; Takens, 1974b) we describe a systematic approach to these resonance phenomena by approximating $\mathbf{f}_\mu^q$, $\mu \in \mathbb{R}$, by the time-2π map of an autonomous vector field. The construction of the approximation is described in § 5.5. This approach has the advantage that it can be extended to include the generic case when $q = 1$ and $q = 2$ (see (ii) above). For these values of q, $\mathbf{Df}_0(\mathbf{0})$ has real eigenvalues and is not a rotation. Bifurcation diagrams for the resulting families of vector fields are presented in § 5.6 and their relation to the corresponding local family $\mathbf{f}|U$ is discussed in § 5.7.

5.2 Arnold's circle map

Consider the following two-parameter family of diffeomorphisms $f_{(\alpha,\varepsilon)}: S^1 \to S^1$,

$$f_{(\alpha,\varepsilon)}(\theta) = \theta + \alpha + \varepsilon \sin \theta, \tag{5.2.1}$$

$\varepsilon \in [0, 1)$, θ, $\alpha \in [0, 2\pi]$, with 0 and 2π identified. For what values of (α, ε) does $f_{(\alpha,\varepsilon)}$ have rotation number p/q in lowest terms, where $p \in \{0, 1, \ldots, q-1\}$, $q \in \mathbb{Z}^+$? It is convenient to measure θ in units of 2π, i.e. $2\pi\theta' = \theta$, so that the iteration $\theta_{n+1} = f_{(\alpha,\varepsilon)}(\theta_n)$ becomes

$$\begin{aligned} \theta'_{n+1} &= \frac{1}{2\pi} f_{(\alpha,\varepsilon)}(2\pi\theta'_n) = f'_{(\alpha',\varepsilon')}(\theta'_n) \\ &= \theta'_n + \alpha' + \varepsilon' \sin 2\pi\theta', \end{aligned} \tag{5.2.2}$$

where $2\pi\alpha' = \alpha$ and $2\pi\varepsilon' = \varepsilon$. Dropping primes, we can write

$$\theta_{n+1} = f_{(\alpha,\varepsilon)}(\theta_n) = \theta_n + \alpha + \varepsilon \sin 2\pi\theta_n, \tag{5.2.3}$$

with α, $\theta \in [0, 1]$ and $\varepsilon \in [0, 1/2\pi)$. Finally, we obtain the lift, $\bar{f}_{(\alpha,\varepsilon)}$, of $f_{(\alpha,\varepsilon)}$ as

$$\bar{f}_{(\alpha,\varepsilon)}(x) = x + \alpha + \varepsilon \sin 2\pi x, \tag{5.2.4}$$

$x \in \mathbb{R}$. This clearly satisfies $\bar{f}_{(\alpha,\varepsilon)}(x+1) = \bar{f}_{(\alpha,\varepsilon)}(x) + 1$, as required for an orientation-preserving diffeomorphism on S^1 (see § 1.2).

Proposition 5.2.1 *The rotation number $\rho(f_{(\alpha,\varepsilon)}) = p/q$ if and only if*

$$\bar{f}^q_{(\alpha,\varepsilon)}(x) - (x + p) = 0 \tag{5.2.5}$$

for some $x \in \mathbb{R}$.

From: Arrowsmith & Place, Dynamical Systems, 1990.

Proof. Clearly, if (5.2.5) is satisfied for some $x = x_0$ then

$$\bar{f}^q_{(\alpha,\varepsilon)}(x_0) = x_0 + p, \tag{5.2.6}$$

and therefore

$$\bar{f}^{nq}_{(\alpha,\varepsilon)}(x_0) = x_0 + np. \tag{5.2.7}$$

Thus

$$\begin{aligned} \rho(f_{(\alpha,\varepsilon)}) &= \operatorname*{Lim}_{n\to\infty} \frac{\bar{f}^n_{(\alpha,\varepsilon)}(x_0) - x_0}{n} \bmod 1 \\ &= \operatorname*{Lim}_{n\to\infty} \frac{\bar{f}^{nq}_{(\alpha,\varepsilon)}(x_0) - x_0}{nq} \bmod 1 = p/q. \end{aligned} \tag{5.2.8}$$

Conversely, observe that (5.2.4) implies

$$\bar{f}^q_{(\alpha,\varepsilon)}(x) = x + q\alpha + F(\alpha, \varepsilon, x). \tag{5.2.9}$$

Let $\alpha = (p/q) + \beta$ to obtain

$$\bar{f}^q_{(\alpha,\varepsilon)}(x) = x + p + G_{p/q}(\beta, \varepsilon, x), \tag{5.2.10}$$

where $G_{p/q}(\beta, \varepsilon, x) = q\beta + F((p/q) + \beta, \varepsilon, x)$. Thus if (5.2.5) is *not* satisfied for some $x \in \mathbb{R}$, then

$$G_{p/q}(\beta, \varepsilon, x) \neq 0 \tag{5.2.11}$$

for *all* $x \in \mathbb{R}$. Since $G_{p/q}$ is periodic in x (see Exercise 5.2.1), this means that $G_{p/q}$ is bounded away from zero. It follows that

$$|\rho(f_{(\alpha,\varepsilon)}) - p/q| \geq \min_{x\in\mathbf{R}} |q^{-1} G_{p/q}(\beta, \varepsilon, x)| > 0. \tag{5.2.12}$$

Hence, $\rho(f_{(\alpha,\varepsilon)}) = p/q$ if and only if (5.2.5) is satisfied for some $x \in \mathbb{R}$. □

It is not difficult to show that

$$F(\alpha, \varepsilon, x) = \varepsilon \sum_{k=0}^{q-1} \sin[2\pi \bar{f}^k_{(\alpha,\varepsilon)}(x)], \tag{5.2.13}$$

and

$$\sin[2\pi \bar{f}^k_{(\alpha,\varepsilon)}(x+1)] = \sin[2\pi \bar{f}^k_{(\alpha,\varepsilon)}(x)], \tag{5.2.14}$$

$k = 0, 1, \ldots, q-1$, so that G is bounded and attains its maximum and minimum value on $[0, 1]$. Thus, for each ε, there is an interval of β on which $G_{p/q}(\beta, \varepsilon, x) = 0$ for some $x \in [0, 1]$ (see Figure 5.2).

How do the end-points of this interval of β depend on ε? For $q = 1$, $p = 0$ and

$$G_{0/1}(\beta, \varepsilon, x) = \beta + \varepsilon \sin 2\pi x = 0 \tag{5.2.15}$$

has solutions for some $x \in [0, 1]$ provided $\beta \leq \pm\varepsilon$. Since 0 and 1 are identified, we

conclude that $\rho(f_{(\alpha,\varepsilon)}) = 0$ for (α, ε) in the linear wedge-shaped regions shown in Figure 5.3(*a*). Given that $q \geqslant 2$, we can approximate the boundary of the region in which $\rho(f_{(\alpha,\varepsilon)}) = p/q$ for $\varepsilon \ll 1$. First observe that (5.2.13) implies

$$|F(\alpha, \varepsilon, x)| \leqslant q\varepsilon \tag{5.2.16}$$

for any α, x. Therefore, $|\beta| \leqslant \varepsilon$ is a necessary condition for $G_{p/q}(\beta, \varepsilon, x)$ to be zero for some x. Thus, if ε is small so is β and we can consider the Taylor expansion for $F((p/q) + \beta, \varepsilon, x)$ about $(\beta, \varepsilon) = (0, 0)$. To this end we next observe that, for $q \geqslant 2$,

$$\bar{f}^k_{(\alpha,\varepsilon)}(x) = \begin{cases} x + k\alpha + \varepsilon \sum_{l=0}^{k-1} \sin(2\pi \bar{f}^l_{(\alpha,\varepsilon)}(x)), & k = 1, \ldots, q-1; \\ x, & k = 0. \end{cases} \tag{5.2.17}$$

Hence, from (5.2.13),

$$\begin{aligned} F(\alpha, \varepsilon, x) &= \varepsilon\left\{\sin 2\pi x + \sum_{k=1}^{q-1} \sin\left[2\pi\left(x + \frac{kp}{q} + k\beta + \varepsilon \sum_{l=0}^{k-1} \sin[2\pi \bar{f}^l_{(\alpha,\varepsilon)}(x)]\right)\right]\right\} \\ &= \varepsilon\left\{\sin 2\pi x + \sum_{k=1}^{q-1} \sin\left[2\pi\left(x + \frac{kp}{q}\right)\right]\right\} + O(\varepsilon^{r+1}\beta^s : r + s = 1), \end{aligned} \tag{5.2.18}$$

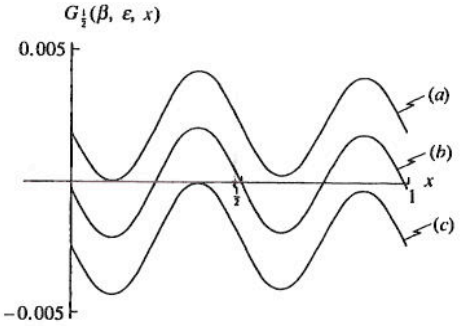
where $\alpha = (p/q) + \beta$. It is easily shown (see Exercise 5.2.2) that

$$\sum_{k=1}^{q-1} \sin\left[2\pi\left(x + \frac{kp}{q}\right)\right] = -\sin 2\pi x; \tag{5.2.19}$$

and we conclude that, for $q \geqslant 2$, $G_{p/q}(\beta, \varepsilon, x)$ takes the form

$$G_{p/q}(\beta, \varepsilon, x) = q\beta + g_0(x)\beta\varepsilon + g_1(x)\varepsilon^2 + O(\varepsilon^{r+1}\beta^s : r + s = 2). \tag{5.2.20}$$

Figure 5.2 Plots of $G_{1/2}(\beta, \varepsilon, x)$ (see (5.2.27)) for $\varepsilon = 0.025$ and (*a*) $\beta = \beta_m(\varepsilon) \simeq 0.001$; (*b*) $\beta = 0$; (*c*) $\beta = -\beta_m(\varepsilon)$. Observe that $G_{1/2}(\beta, \varepsilon, x) = 0$ for some $x \in [0, 1]$ provided $\beta \in [-\beta_m(\varepsilon), \beta_m(\varepsilon)]$.



Let $G_{p/q}(\beta, \varepsilon, x) = 0$ define β as a function of ε and x in the form

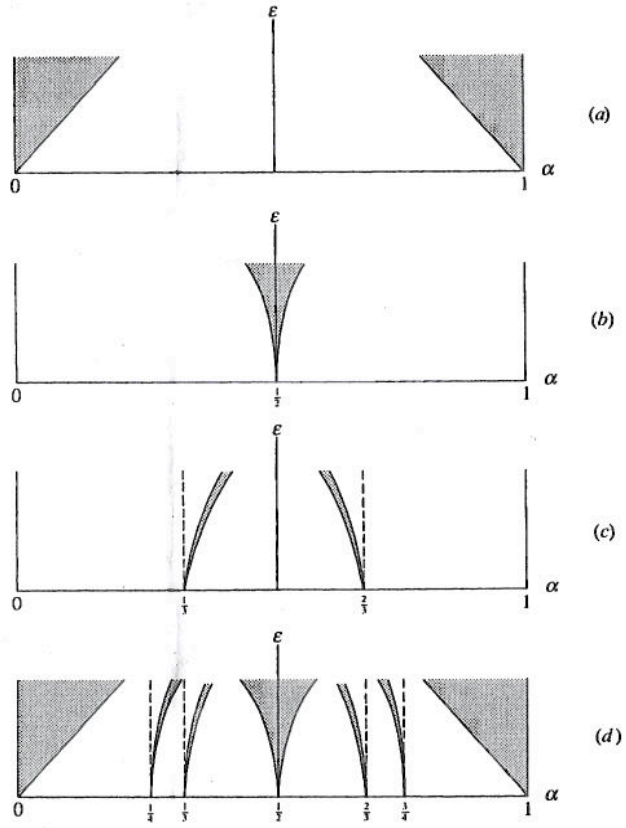
$$\beta(\varepsilon, x) = \beta_0(x) + \varepsilon\beta_1(x) + \varepsilon^2\beta_2(x) + O(\varepsilon^3). \tag{5.2.21}$$

Substitute (5.2.21) into (5.2.20) and compare coefficients of powers of ε to obtain

$$\beta_0(x) = \beta_1(x) \equiv 0, \qquad \beta_2(x) = -\frac{g_1(x)}{q}. \tag{5.2.22}$$

It follows that $\beta(\varepsilon, x)$ is at least quadratic in ε for all $q \geqslant 2$.

Figure 5.3 Arnold tongues for the circle map (5.2.3): (*a*) $T_{0/1}$; (*b*) $T_{1/2}$; (*c*) $T_{1/3}$ and $T_{2/3}$; (*d*) schematic illustration of $\{T_{p/q} | 1 \leqslant q \leqslant 4\}$. There is a tongue $T_{p/q} = \{(\alpha, \varepsilon) | \rho(f_{(\alpha,\varepsilon)}) = p/q\}$ for every rational number $p/q \in [0, 1)$ but its width decreases rapidly with increasing q. The system of tongues is symmetric about $\alpha = \frac{1}{2}$.



Example 5.2.1 Given that $p/q = \frac{1}{2}$, find $g_1(x)$. Describe the set $\{(\alpha, \varepsilon)|\rho(f_{(\alpha,\varepsilon)}) = \frac{1}{2}\}$ and draw a diagram illustrating this set in the α,ε-plane.

Solution. Observe that

$$\bar{f}^2_{(\alpha,\varepsilon)}(x) = x + 2\alpha + \varepsilon \sin 2\pi x + \varepsilon \sin 2\pi(x + \alpha + \varepsilon \sin 2\pi x). \quad (5.2.23)$$

Let $\alpha = \frac{1}{2} + \beta$ to obtain

$$\bar{f}^2_{(\alpha,\varepsilon)}(x) = x + 1 + 2\beta + \varepsilon \sin 2\pi x + \varepsilon \sin 2\pi(x + \tfrac{1}{2} + \beta + \varepsilon \sin 2\pi x). \quad (5.2.24)$$

Therefore

$$G_{1/2}(\beta, \varepsilon, x) = 2\beta + \varepsilon \sin 2\pi x + \varepsilon \sin 2\pi(x + \tfrac{1}{2} + \beta + \varepsilon \sin 2\pi x). \quad (5.2.25)$$

Now,

$$\varepsilon \sin 2\pi(x + \tfrac{1}{2} + \beta + \varepsilon \sin 2\pi x) = \varepsilon\{\sin(2\pi(x + \tfrac{1}{2})) \cos(2\pi(\beta + \varepsilon \sin 2\pi x)) + \cos(2\pi(x + \tfrac{1}{2})) \sin(2\pi(\beta + \varepsilon \sin 2\pi x))\} \quad (5.2.26)$$

and

$$G_{1/2}(\beta, \varepsilon, x) = 2\beta + \varepsilon \sin 2\pi x + \varepsilon\{\sin(2\pi(x + \tfrac{1}{2})) + \cos(2\pi(x + \tfrac{1}{2}))[2\pi(\beta + \varepsilon \sin 2\pi x)]\} + O(\varepsilon^{r+1}\beta^s : r + s = 2). \quad (5.2.27)$$

Thus

$$g_1(x) = 2\pi \cos(2\pi(x + \tfrac{1}{2})) \sin 2\pi x, = -2\pi \cos 2\pi x \sin 2\pi x. \quad (5.2.28)$$

Therefore

$$\beta(\varepsilon, x) = \frac{\pi\varepsilon^2}{2} \sin 4\pi x + O(\varepsilon^3). \quad (5.2.29)$$

For given $\varepsilon \ll 1$,

$$-\frac{\pi}{2}\varepsilon^2 + O(\varepsilon^3) \leqslant \beta(\varepsilon, x) \leqslant \frac{\pi\varepsilon^2}{2} + O(\varepsilon^3) \quad (5.2.30)$$

and $\{(\alpha, \varepsilon)|\rho(f_{(\alpha,\varepsilon)}) = \frac{1}{2}\}$ is a symmetric cusp-shaped region with vertex at $\alpha = \frac{1}{2}$ (see Figure 5.3(*b*)). □

The sets $T_{p/q} = \{(\alpha, \varepsilon)|\rho(f_{(\alpha,\varepsilon)}) = p/q\}$ for $p/q \in \mathbb{Q} \cap [0, 1)$ are known as 'Arnold tongues' and their boundaries can be approximated for any p/q (see Exercise 5.2.3). For example, when $p/q = \frac{1}{3}$, it can be shown that $\beta_2(x) \equiv 3^{1/2}\pi/6$. This means that the approximations to the upper and lower boundaries of $T_{1/3}$ have the same quadratic terms and the tongue must leave $\alpha = \frac{1}{3}$ as shown in Figure 5.3(*c*). The width of $T_{1/3}$ is at most $O(\varepsilon^3)$. Such asymmetric behaviour is typical of tongues

with $q \geqslant 3$ and the symmetry exhibited by $T_{0/1}$ and $T_{1/2}$ is exceptional. It is also useful to note that the system of tongues is symmetric about $\alpha = \frac{1}{2}$. More precisely, $T_{(1-p/q)}$ is the image of $T_{p/q}$ under reflection in $\alpha = \frac{1}{2}$. This means that $T_{2/3}$ takes the form shown in Figure 5.3(*c*). Similar considerations for $q = 4$ yield the schematic representation of $\{T_{p/q}|1 \leqslant q \leqslant 4\}$ given in Figure 5.3*d*.

There is a separate tongue for every rational in $[0, 1)$. The greater the value of q the thinner is the tongue, but each one still has a positive width for $\varepsilon > 0$. It follows that the dependence of $\rho(f_{(\alpha,\varepsilon)})$ on α, for fixed $\varepsilon > 0$, is rather subtle. For each rational number, $p/q \in [0, 1)$, there is an interval of values of α for which $\rho(f_{(\alpha,\varepsilon)}) = p/q$. However, the length of the (p/q)-interval diminishes rapidly with increasing q and, in fact, the measure of the totality of tongues for $0 < \varepsilon < \varepsilon_0 \ll (2\pi)^{-1}$, $0 \leqslant \alpha \leqslant 1$ is small compared with ε_0. This means that, on selecting a member of the family at random, it will have an irrational rotation number with probability near one as $\varepsilon \to 0$. This is in sharp contrast to the behaviour of circle diffeomorphisms for which (see Theorem 3.4.1) a rational rotation number is a generic property.

Analogous results to those described above can be shown to hold for any analytic or sufficiently smooth unfolding of a rotation. In particular, this is true for families of the form

$$f_{(\alpha,\varepsilon)}(\theta) = \theta + \alpha + \varepsilon a(\theta), \quad (5.2.31)$$

where $a(\theta)$ is an arbitrary analytic function on S^1.

5.3 Irrational rotations

When $\mathrm{Df}_0(\mathbf{0})$ is an irrational rotation, a straightforward normal form calculation shows that an invariant circle occurs. We have already considered this calculation for $\mu = 0$ in Examples 2.5.2 and 3, where it was shown that f_0 could be transformed into the complex form

$$\tilde{f}_0(z) = \lambda(0)z + a_3 z|z|^2 + O(|z|^5), \quad (5.3.1)$$

where $a_3 = \tilde{a}_{21} \in \mathbb{C}$ in (2.5.31) and $\lambda(0) = \exp(2\pi i\beta)$. Let $\lambda(\mu), \bar{\lambda}(\mu)$ be the eigenvalues of $\mathrm{Df}_\mu(\mathbf{0})$. Then, given that $[(\mathrm{d}|\lambda(\mu)|/\mathrm{d}\mu)]_{\mu=0} \neq 0$, $|\lambda(\mu)| \neq 1$ for $\mu \neq 0$ and the $z|z|^2$-term in (5.3.1) is no longer resonant. Therefore, it could be removed by a suitable transformation. However, this would destroy the continuity of $\bar{f}(\mu, z)$ in μ. We therefore choose not to remove this term and conclude that $f(\mu, z)$ is equivalent to

$$\tilde{f}_\mu(z) = \lambda(\mu)z\{1 + a(\mu)|z|^2 + R(\mu, z)\}, \quad (5.3.2)$$

where $\lambda(\mu) = |\lambda(\mu)|\exp(2\pi i\beta(\mu))$ and $a(\mu) = a_3(\mu)/\lambda(\mu)$ depends smoothly on μ. In (5.3.2), $R(\mu, z)$ is $O(|z|^4)$.

Let $a(\mu) = c(\mu) + id(\mu)$, then, in the absence of the remainder, $R(\mu, z)$,

$$|\tilde{f}_\mu(z)| = |\lambda(\mu)|\,|z|\,|(1 + a(\mu)|z|^2)|. \quad (5.3.3)$$