

4.4 ITERATED FUNCTION SYSTEMS

*New ways of looking at something
can have surprising and useful
applications.*

Overview

This Section corresponds to the second half of Section 6.3 of [HM], pp 437–450]. The material is developed here in much more detail.

We discuss the idea of an *Iterated Function System* or *IFS*, a very useful way of examining a large class of fractals:

In particular, the idea of an IFS leads to two different ways of generating fractals, the *Deterministic Algorithm* and the *Random Algorithm* or *Chaos Game*.

We discuss the IFS corresponding to the Sierpinski Triangle and the IFS corresponding to the Koch curve. But the ideas in these two cases can be generalised in a more or less straightforward way to any IFS.

A Little History The idea that many fractals can be characterised by an IFS and that such fractals can be generated by the deterministic algorithm was introduced and developed in a 1981 paper¹⁰ of the author.

The idea of the chaos game to generate fractals was first developed by Barnsley and Demko¹¹ in 1985. A few years later Barnsley applied these ideas to image compression and was a founder of the company “Iterated Systems”, at one stage valued at \$200,000,000(US), later known as “Media Bin” and then acquired by “Interwoven”.

What is an IFS?

[HM, 437–446]

Three Maps and the Sierpinski Triangle S If you look at the Sierpinski Triangle S in Figure 4.20 you can see that S is made up of three copies of itself each scaled by $1/2$.¹² We will denote these three copies by S_1 , S_2 and S_3 , where the vertex $P_1 = (0, 0) \in S_1$, the vertex $P_2 = (1, 0) \in S_2$ and the top vertex $P_3 = (1/2, \sqrt{3}/2) \in S_3$. *Indicate all this on Figure 4.20.*

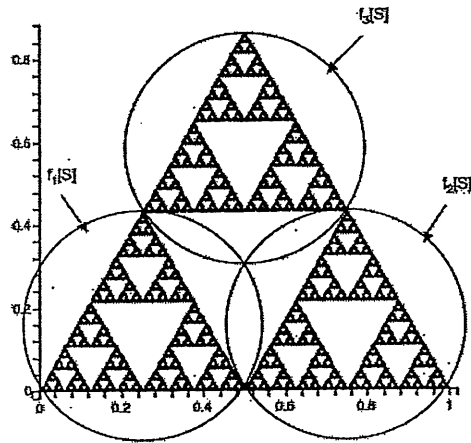
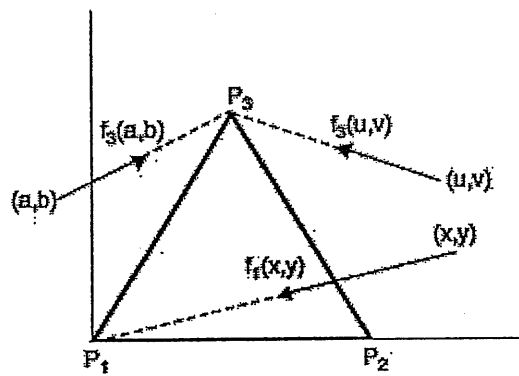
Suppose that each point (x, y) in the plane \mathbb{R}^2 is moved closer to the point P_1 by the factor $1/2$. For example, $(3, 0)$ is mapped to $(3/2, 0)$, $(1, 1)$ is mapped to $(1/2, 1/2)$, $(0, 3)$ is mapped to $(0, 3/2)$.

In this way, points in S are mapped to points in S . For example, $(1, 0)$ is mapped to $(0, 0.5)$, $(3/4, \sqrt{3}/4)$ which is the midpoint of the “right edge” of S is mapped to $(3/8, \sqrt{3}/8)$ (*where is this on S ?*) and $P_3 = (1/2, \sqrt{3}/2)$ is mapped to $(1/4, \sqrt{3}/4)$. (*where is this on S ?*)

¹⁰Hutchinson, John E., *Fractals and self-similarity*. Indiana Univ. Math. J. 30 (1981), 713–747.

¹¹Barnsley, M. F.; Demko, S. *Iterated function systems and the global construction of fractals*. Proc. Roy. Soc. London Ser. A 399 (1985), 243–275.

¹²Of course, as usual, we can only sketch an approximation to S .

Figure 4.20: Sierpinski Triangle S Figure 4.21: The maps f_1 , f_2 and f_3 .

The map (or function) we just described is the function $f_1 : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined by

$$f_1(x, y) = \left(\frac{x}{2}, \frac{y}{2} \right), \quad (4.11)$$

see Figure 4.21.

Notice from Figure 4.20 that

$$S_1 = f_1[S],$$

where by the right side of the equality we mean the set of all points of the form $f_1(x, y)$ for $(x, y) \in S$. That is

$$f_1[S] = \{f_1(x, y) : (x, y) \in S\}.$$

We read this as " $f_1[S]$ is the set of points of the form $f_1(x, y)$ for some $(x, y) \in S$ ".

In a similar way,

$$S_2 = f_2[S] \quad \text{and} \quad S_3 = f_3[S],$$

where f_2 maps every point $(x, y) \in \mathbb{R}^2$ to the midpoint between (x, y) and P_2 , and f_3 maps every point $(x, y) \in \mathbb{R}^2$ to the midpoint between (x, y) and P_3 .

Why are the following formulae true?

$$\begin{aligned} f_1(x, y) &= \left(\frac{x}{2}, \frac{y}{2} \right) \\ f_2(x, y) &= (1, 0) + \frac{1}{2} \left((x, y) - (1, 0) \right) = \left(\frac{x}{2} + \frac{1}{2}, \frac{y}{2} \right) \\ f_3(x, y) &= \left(\frac{1}{2}, \frac{\sqrt{3}}{2} \right) + \frac{1}{2} \left((x, y) - \left(\frac{1}{2}, \frac{\sqrt{3}}{2} \right) \right) = \left(\frac{x}{2} + \frac{1}{4}, \frac{y}{2} + \frac{\sqrt{3}}{4} \right). \end{aligned} \quad (4.12)$$

If you know about matrices and column vectors, then f_1 , f_2 and f_3 can also be conveniently described that way.¹³

The IFS for S We have seen that for the Sierpinski triangle S ,

$$S = S_1 \cup S_2 \cup S_3$$

where

$$S_1 = f_1[S], \quad S_2 = f_2[S], \quad S_3 = f_3[S].$$

So we have the important relation

$$S = f_1[S] \cup f_2[S] \cup f_3[S] \quad (4.13)$$

Definition 4.4.1. The set of maps $\mathcal{F} = \{f_1, f_2, f_3\}$ with f_1 , f_2 and f_3 as in (4.12) is called the *Iterated Function System* or *IFS* corresponding to the Sierpinski Triangle S .

Because of (4.13) we say that S is *invariant* under the IFS $\mathcal{F} = \{f_1, f_2, f_3\}$.

¹³If we represent points in the plane by column vectors then it follows from (4.12) that

$$f_1 \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \quad f_2 \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \frac{1}{2} \\ 0 \end{bmatrix}, \quad f_3 \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \frac{1}{4} \\ \frac{\sqrt{3}}{4} \end{bmatrix}.$$

Contractive Maps Take two initial points $x = (x, y)$ and $y = (u, v)$ and apply the map f_1 to each. The image points are $f_1(x, y) = (x/2, y/2)$ and $f_1(u, v) = (u/2, v/2)$ respectively, see (4.11).

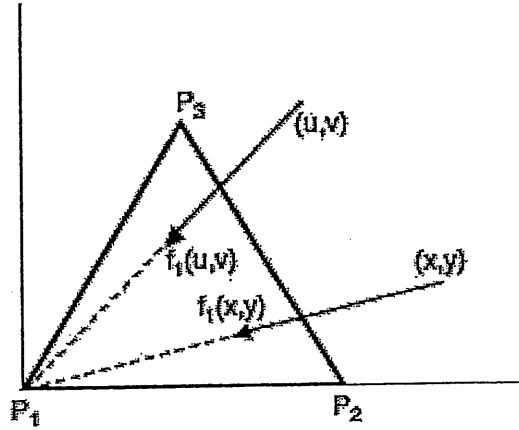


Figure 4.22: f_1 is contractive

The distance¹⁴ between the two image points is exactly $1/2$ the distance between the two initial points. This is essentially because of the factor $1/2$ in the definition of f_1 , see (4.11).

If you know about vectors, then you will see that the vector from (x, y) to (u, v) can be written as $(u - x, v - y)$. The vector from $(x/2, y/2)$ to $(u/2, v/2)$ is $(u/2 - x/2, v/2 - y/2) = \frac{1}{2}(u - x, v - y)$, which is exactly half the length of the vector $(u - x, v - y)$ from (x, y) to (u, v) .

You could also use the formula for computing the distance d between points. For example,

$$\begin{aligned} d(f_1(x, y), f_1(u, v)) &= d((x/2, y/2), (u/2, v/2)) \\ &= \sqrt{(x/2 - u/2)^2 + (y/2 - v/2)^2} \\ &= \frac{1}{2} \sqrt{(x - u)^2 + (y - v)^2} \\ &= \frac{1}{2} d((x, y), (u, v)) \end{aligned} \quad (4.14)$$

In the same way we see that the maps f_2 and f_3 also reduce the distance between points by the factor $1/2$.

Definition 4.4.2. A function $f : \mathbb{R} \rightarrow \mathbb{R}$, $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ or $f : \mathbb{R}^3 \rightarrow \mathbb{R}^3$, is *contractive* if there is a number r with $0 \leq r < 1$ such that

$$d(f(x), f(u)) \leq r d(x, u)$$

¹⁴The distance between $x, y \in \mathbb{R}$ is $d(x, y) = |x - y|$.
The distance between $x = (x_1, x_2)$, $y = (y_1, y_2) \in \mathbb{R}^2$ is $d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$.
The distance between $x = (x_1, x_2, x_3)$, $y = (y_1, y_2, y_3) \in \mathbb{R}^3$ is $d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$.

for all points x and u in \mathbb{R} , \mathbb{R}^2 or \mathbb{R}^3 respectively.

The number r is called a *contractivity factor*¹⁵ for f .

For example, it follows from (4.14) that f_1 is contractive with contractivity factor $1/2$. Similarly, f_2 and f_3 are also contractive with contractivity factor $1/2$.

It will generally be the case, at least in this course, that the maps in an IFS are all contractive.

The Deterministic Algorithm

[HM, 437–446]

The IFS Determines S A surprising and very important fact is that from just knowing the IFS $\mathcal{F} = \{f_1, f_2, f_3\}$ in Definition 4.4.1 we can find S .

To see this, begin with *any* set (picture) E , such as the face in Figure 4.23, and apply the IFS \mathcal{F} to get a new set (picture)

$$E_1 = \mathcal{F}(E) = f_1[E] \cup f_2[E] \cup f_3[E].$$

So E_1 consists of 3 little faces.

Next apply \mathcal{F} to E_1 to get a new set (picture) E_2 consisting of 9 smaller faces.

$$E_2 = \mathcal{F}(E_1) = f_1[E_1] \cup f_2[E_1] \cup f_3[E_1].$$

Next apply \mathcal{F} to E_2 to get a new set (picture) E_3 consisting of 27 smaller faces.

$$E_3 = \mathcal{F}(E_2) = f_1[E_2] \cup f_2[E_2] \cup f_3[E_2].$$

And so on.

In the limit we obtain the Sierpinski Triangle S , no matter what set E we start from. See Theorem 4.4.3.

For any set E we defined the set

$$\mathcal{F}(E) = f_1[E] \cup f_2[E] \cup f_3[E]. \quad (4.15)$$

We have the following result.


Theorem 4.4.3. Consider the IFS $\mathcal{F} = \{f_1, f_2, f_3\}$ in Definition 4.4.1. Then the sequence of sets

$$E, E_1 = \mathcal{F}(E), E_2 = \mathcal{F}(E_1), E_3 = \mathcal{F}(E_2), \dots, E_n = \mathcal{F}(E_{n-1}), \dots \quad (4.16)$$

converges to S , independent of the starting set E , provided E is bounded.¹⁶

¹⁵Notice that if r is a contractivity factor then so is any number larger than r . One can show there is always a smallest contractivity factor, and then this is usually called the contractivity factor for f .

¹⁶A set $E \subset \mathbb{R}, \mathbb{R}^2, \mathbb{R}^3$ is *bounded* if there is some number M such that the distance from every point in E to the origin is at most M . For example, the Sierpinski triangle and the Koch curve and the set of points inside any disc, are all bounded. But the entire plane \mathbb{R}^2 is not bounded, nor is the set of all points (x, y) in the plane for which $x \geq 0$ and $y \geq 0$.

 If we take $E = \mathbb{R}^2$ in the statement of the Theorem, then every set in the sequence is \mathbb{R}^2 . Why? So the sequence of sets in this case will not converge to S .

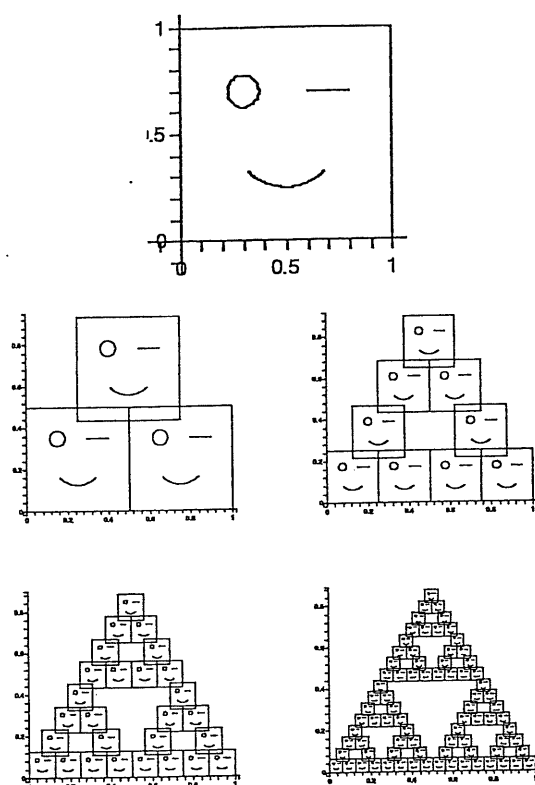


Figure 4.23: A sequence of sets $E, E_1, E_2, E_3, E_4, \dots$, beginning with a face and obtained by repeatedly applying the IFS $\mathcal{F} = \{f_1, f_2, f_3\}$, converging to the Sierpinski Triangle S .

“Proof”. We cannot give a complete and rigorous proof, as we have not defined what we mean by the limit of a sequence of sets.

Also, to make the following rigorous requires the filling in of quite a few details about converging sequences of sets.

But I will describe the essential parts of the argument.

Somewhat informally, by saying the set S is the limit of the sequence of sets (4.16), i.e. the sequence of sets converges to S , we mean that for *every* positive number ϵ ,¹⁷ which we can think of as a very small “tolerance”, the following is true:

- There is an integer N depending on ϵ such that if $n \geq N$ then
1. for every $x \in E_n$ there is some $y \in S$ within distance ϵ of x ,¹⁸ and
2. for every $y \in S$ there is some $x \in E_n$ within distance ϵ of y .¹⁹

¹⁷In Mathematics we commonly use the Greek letter ϵ , called “epsilon”, to represent a number which is small and positive.

¹⁸The point y depends on x , on ϵ and on n .

¹⁹The point x depends on y , on ϵ and on n .

From the way we originally defined the Sierpinski Triangle S beginning on page 151, we know S is the limit of the sequence

$$T, T_1 = \mathcal{F}(T), T_2 = \mathcal{F}(T_1), T_3 = \mathcal{F}(T_2), \dots, T_n = \mathcal{F}(T_{n-1}), \dots, \quad (4.17)$$

where T is the black equilateral triangle as in Figure 4.24.

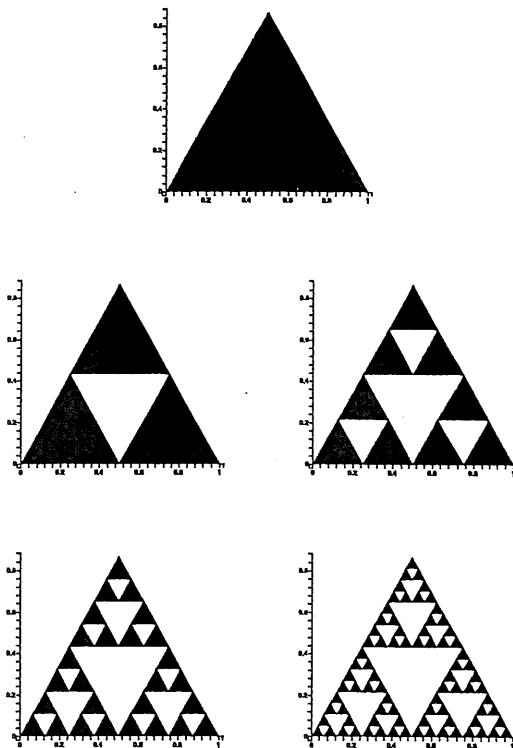


Figure 4.24: A sequence of sets $T, T_1, T_2, T_3, T_4, \dots$, beginning with a triangle and obtained by repeatedly applying the IFS $\mathcal{F} = \{f_1, f_2, f_3\}$, converging to the Sierpinski Triangle S .

There is certainly some positive real number, let us call it α , such that

1. every point in the triangle T is within distance α of some point in the set E , and
2. every point in the set E is within distance α of some point in the triangle T .

For example, $\alpha = 1$ would do if E is the face in Figure 4.23. In fact smaller α will also work, but that does not make a difference to the following proof. *What is a smaller α that works?*

Because the functions f_1, f_2 and f_3 contract distances by $1/2$,

1. every point in T_1 is within distance $\alpha/2$ of some point in E_1 , and

2. every point in E_1 is within distance $\alpha/2$ of some point in T_1 .

Again because the functions f_1 , f_2 and f_3 contract distances by $1/2$,

1. every point in T_2 is within distance $\alpha/4$ of some point in E_2 , and
2. every point in E_2 is within distance $\alpha/4$ of some point in T_2 .

Again because the functions f_1 , f_2 and f_3 contract distances by $1/2$,

1. every point in T_3 is within distance $\alpha/8$ of some point in E_3 , and
2. every point in E_3 is within distance $\alpha/8$ of some point in T_3 .

Etc.

Beginning on page 151 we essentially saw that the sequence (4.17) converges to S , in fact this was essentially how we defined S . We also have just seen that the sets in the sequence (4.16) are getting closer and closer to the sets in the sequence (4.17). It follows that the sets in the sequence (4.16) also converge to S .

This argument did not depend on the initial set E . For different E we will get a different α , but nothing else changes. \square

Where in the proof did we use the fact that E was bounded?

Deterministic Algorithm for Generating S This is what we have just discussed. Begin with any set E and take the sequence (4.16). This will give better and better approximations to S .

The terminology “deterministic algorithm” is used to distinguish this algorithm from the “random algorithm” or “chaos game” discussed on page 176.

There is a nice java applet at www.geom.uiuc.edu/java/IFSoft/IFSS/welcome.html#findingattractors. Scroll down to the blue window, draw your own face, and use the Iterate button to step through successive iterations. Note that the Sierpinski Triangle there, and hence the three functions used, are a little different from the example we have just been discussing.

The Koch Curve K and its IFS The Koch Curve K and its first approximation is shown in Figure 4.25. See also page 150.

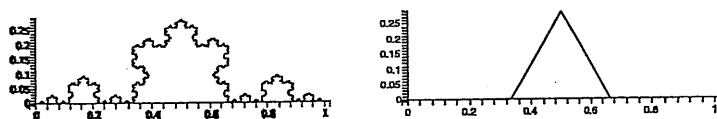


Figure 4.25: Koch Curve and its first approximation

The four line segments in the first approximation each have length $1/3$. The second and third line segments form an equilateral triangle with the x -axis. From this it is easy to check that the vertices of the five corners are



$P_1 = (0,0)$, $P_2 = (1/3,0)$, $Q = (1/2, \sqrt{3}/6)$, $P_3 = (2/3,0)$ and $P_4 = (1,0)$.
Check it!

The Koch Curve K can be written as the union of 4 scaled copies of itself each scaled by $1/3$.

$$K = K_1 \cup K_2 \cup K_3 \cup K_4 = f_1[K] \cup f_2[K] \cup f_3[K] \cup f_4[K]. \quad (4.18)$$

Here K_1 is the left "quarter" joining the points $(0,0)$ and $(0,1/3)$, K_2 joins $(0,1/3)$ and $(1/2, \sqrt{3}/6)$, K_3 joins $(1/2, \sqrt{3}/6)$ and $(2/3,0)$, K_4 joins $(2/3,0)$ and $(1,0)$.

Geometrically:

1. f_1 contracts points towards $(0,0)$ with contraction ratio $1/3$;
2. f_2 contracts points towards $(0,0)$ with contraction ratio $1/3$, then rotates anti clockwise, i.e. in the "positive" direction, by 60° or equivalently $\pi/3$ radians, and finally translates in the x -direction by $1/3$;
3. f_3 contracts points towards $(0,0)$ with contraction ratio $1/3$, then rotates by -60° or equivalently $-\pi/3$ radians, and finally translates the origin $(0,0)$ to $Q = (1/2, \sqrt{3}/6)$;
4. f_4 contracts points towards $(0,0)$ with contraction ratio $1/3$ and then translates in the x -direction by $2/3$.

The formulae for f_1, \dots, f_4 are:

$$\begin{aligned} f_1(x,y) &= (x/3, y/3), \\ f_2(x,y) &= (x/6 - \sqrt{3}y/6 + 1/3, \sqrt{3}x/6 + y/6), \\ f_3(x,y) &= (x/6 + \sqrt{3}y/6 + 1/2, -\sqrt{3}x/6 + y/6 + \sqrt{3}/6), \\ f_4(x,y) &= (x/3 + 2/3, y/3). \end{aligned} \quad (4.19)$$

If you know a little about matrices and how to represent rotations by matrices, the geometric descriptions will allow you to compute the functions f_1, \dots, f_4 .²⁰

The IFS corresponding to K is

$$\mathcal{F} = \{f_1, f_2, f_3, f_4\}, \quad (4.20)$$

where f_1, \dots, f_4 are as in (4.19).

The contractivity factor for the maps f_1, \dots, f_4 is $1/3$. Why?

If E is any set then we define

$$\mathcal{F}(E) = f_1[E] \cup f_2[E] \cup f_3[E] \cup f_4[E].$$

Just as in Theorem 4.4.3 for the Sierpinski Triangle, there is a similar theorem for the Koch Curve.

²⁰For example, from the description of f_2 , the point (x,y) is first sent to $(x/3, y/3)$. Since a rotation by θ radians is represented by the matrix $\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$, it follows that $(x/3, y/3)$ is then sent to

$$\begin{bmatrix} 1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & 1/2 \end{bmatrix} \begin{bmatrix} x/3 \\ y/3 \end{bmatrix} = \begin{bmatrix} x/6 - \sqrt{3}y/6 \\ \sqrt{3}x/6 + y/6 \end{bmatrix}.$$

Finally, translation by $1/3$ in the x -direction adds $1/3$ to the first coordinate. This gives the formula for $f_2(x,y)$.



Use a similar argument to find the formulae for $f_4(x,y)$ and $f_3(x,y)$.

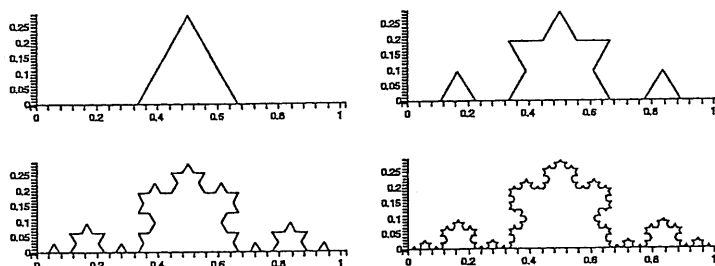


Figure 4.26: A sequence of sets which began with a line segment, obtained by repeatedly applying the IFS $\mathcal{F} = \{f_1, f_2, f_3, f_4\}$ in (4.20), converging to the Koch Curve K .

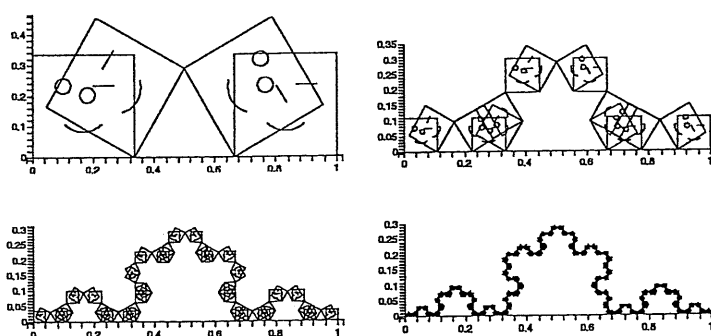


Figure 4.27: A sequence of sets which began with a face, obtained by repeatedly applying the IFS $\mathcal{F} = \{f_1, f_2, f_3, f_4\}$, converging to the Koch Curve K .

Theorem 4.4.4. Consider the IFS $\mathcal{F} = \{f_1, f_2, f_3, f_4\}$ in (4.20). Then the sequence of sets

$$E, E_1 = \mathcal{F}(E), E_2 = \mathcal{F}(E_1), E_3 = \mathcal{F}(E_2), \dots, E_n = \mathcal{F}(E_{n-1}), \dots \quad (4.21)$$

converges to K , independent of the starting set E , provided E is bounded.

“Proof”. The proof is very similar to that for the Sierpinski Triangle. The only significant difference is that the contractivity factor here for the maps f_1, \dots, f_4 is $1/3$ instead of $1/2$, as was the case for the Sierpinski Triangle IFS. \square

The General IFS Theorem We have the following very general result, Theorem 4.4.5. It extends Theorems 4.4.3 and 4.4.4.

In order to give a precise statement we need the idea of a *closed* and *bounded* set. We have already seen in Footnote 16 what it means for a set to be bounded.

A set E is said to be *closed* if it contains all its boundary points.

3.1 Similarity and Scaling

Self-similarity extends one of the most fruitful notions of elementary geometry: similarity. Two objects are similar if they have the same shape, regardless of their size. Corresponding angles, however, must be equal, and corresponding line segments must all have the same factor of proportionality. For example, when a photo is enlarged, it is enlarged by the same factor in both horizontal and vertical directions. Even an oblique, i.e. non-horizontal, non-vertical, line segment between two points on the original will be enlarged by the same factor. We call this enlargement factor *scaling factor*. The transformation between the objects is called similarity transformation.

What is Similarity?

Similarity Transformations

Similarity transformations are compositions involving a scaling, a rotation and a translation. A reflection may additionally be included, but we skip the details of that. Let us be more specific for similarity transformations in the plane. Here we denote points P by their coordinate pairs $P = (x, y)$. Let us apply scaling, rotation and translation to one point $P = (x, y)$ of a figure. First, a scaling operation, denoted by S , takes place yielding a new point $P' = (x', y')$. In formulas,

$$\begin{aligned}x' &= sx, \\y' &= sy,\end{aligned}$$

where $s > 0$ is the scale factor. A scale reduction occurs, if $s < 1$, and an enlargement of the object will be produced when $s > 1$. Next, a rotation R is applied to $P' = (x', y')$ yielding $P'' = (x'', y'')$.

$$\begin{aligned}x'' &= \cos \theta \cdot x' - \sin \theta \cdot y', \\y'' &= \sin \theta \cdot x' + \cos \theta \cdot y' .\end{aligned}$$

This describes a counterclockwise (mathematically positive) rotation of P' about the origin of the coordinate system by an angle of θ . Finally, a translation T of P'' by a displacement (T_x, T_y) is given by

$$\begin{aligned}x''' &= x'' + T_x, \\y''' &= y'' + T_y\end{aligned}$$

which yields the point $P''' = (x''', y''')$. Summarizing, we may write

$$P''' = T(P'') = T(R(P')) = T(R(S(P)))$$

or, using the notation

$$W(P) = T(R(S(P)))$$

we have $P''' = W(P)$. W is the similarity transformation. In formulas,

$$\begin{aligned}x''' &= s \cos \theta \cdot x - s \sin \theta \cdot y + T_x, \\y''' &= s \sin \theta \cdot x + s \cos \theta \cdot y + T_y .\end{aligned}$$

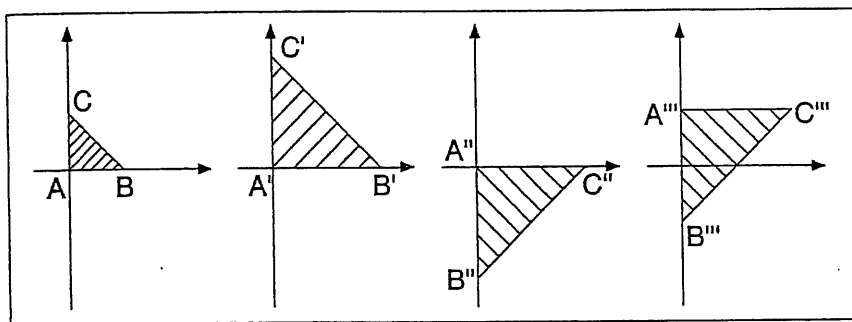


Figure 3.2 : A similarity transformation is applied to the triangle ABC . The scaling factor is $s = 2$, the rotation is by 270° , and the translation is given by $T_x = 0$ and $T_y = 1$.

Applying W to all points of an object in the plane produces a figure which is similar to the original.

The similarity transformations can also be formulated mathematically for objects in other dimensions, for example for shapes in three or only one dimension. In the latter case we have points x on the real line, and the similarity transformation can simply be written as $W(x) = sx + t$, $s \neq 0$.

Consider a photo which is enlarged by a factor of 3. Note that the area of the resulting image is $3 \cdot 3 = 3^2 = 9$ times the area of the original. More generally, if we have an object with area A and scaling factor s , then the resulting object will have an area which is $s \cdot s = s^2$ times the area A of the original. In other words, the area of the scaled-up object increases as the square of the scaling factor.

Scaling 3D-Objects

What about scaling three-dimensional objects? If we take a cube and enlarge it by a scaling factor of 3, it becomes 3 times as long, 3 times as deep, and 3 times as high as the original. We observe that the area of each face of the enlarged cube is $3^2 = 9$ times as large as the face of the original cube. Since this is true for all six faces of the cube, the total surface area of the enlargement is 9 times as much as the original. More generally, for objects of any shape whatever, the total surface area of a scaled-up object increases as the square of the scaling factor.

What about volume? The enlarged cube has 3 layers, each with $3 \cdot 3 = 9$ little cubes. Thus the total volume is $3 \cdot 3 \cdot 3 = 3^3 = 27$ times as much as the original cube. In general, the volume of a scaled-up object increases as the cube of the scaling factor.

These elementary observations have remarkable consequences, which were the object of discussion by Galileo (1564–1642) in his 1638 publication *Dialogues Concerning Two New Sciences*. In fact Galileo² suggested 300

²We quote D'Arcy Thompson's account from his famous 1917 *On Growth and Form* (New Edition, Cambridge University Press, 1942, page 27): "[Galileo] said that if we tried building ships, palaces or temples of enormous size, yards, beams and

Number k	Scale s	Rotation θ	Translation	
			T_x	T_y
1	1/3	0°	0	0
2	1/3	60°	1/3	0
3	1/3	-60°	1/2	$\sqrt{3}/6$
4	1/3	0°	2/3	0

Table 3.20 : Similarity transformations of the Koch curve collage. The transformations are carried out first by applying the scaling, then the rotation, and finally the translation (see section 3.1).

Transformation	x -Part	y -Part
$w_1(x, y)$	$\frac{1}{3}x$	$\frac{1}{3}y$
$w_2(x, y)$	$\frac{1}{6}x - \frac{\sqrt{3}}{6}y + \frac{1}{3}$	$\frac{\sqrt{3}}{6}x + \frac{1}{6}y$
$w_3(x, y)$	$\frac{1}{6}x + \frac{\sqrt{3}}{6}y + \frac{1}{2}$	$-\frac{\sqrt{3}}{6}x + \frac{1}{6}y + \frac{\sqrt{3}}{6}$
$w_4(x, y)$	$\frac{1}{3}x + \frac{2}{3}$	$\frac{1}{3}y$

Table 3.21 : Explicit formulas for the similarity transformations of the Koch curve collage.

We obtain explicit formulas for the transformations as given in table 3.21.

Characterization by an Equation for the Self-Similarity

This collage-like operation can be described by a single mathematical transformation. We let w_1, \dots, w_4 be the four similarity transformations given by a reduction with factor 1/3 composed with a positioning (rotation and translation) at piece k along the polygon as shown in figure 3.22 (bottom). Then, if A is any image, let $W(A)$ denote the collection (union) of all four transformed copies

$$W(A) = w_1(A) \cup w_2(A) \cup w_3(A) \cup w_4(A). \quad (3.9)$$

This is a transformation of images, or more precisely, subsets of the plane. Figure 3.23 shows the result of this transformation when applied to an arbitrary image, for example, when A is the word 'KÖCH'. When comparing the results in figure 3.22 and figure 3.23, we make a fundamental observation. In the case where we apply the transformation W from eqn. (3.9) to the image of the Koch curve, we obtain the Koch curve back again. That is, if we formally introduce a symbol K for the Koch curve, we have the important identity

$$W(K) = K,$$

which is the desired invariance (or fixed-point) property. In other words, if we pose the problem of finding a solution X to the equation $W(X) = X$,

The Koch Collage

The Koch curve is invariant under the transformations w_1 to w_4 .

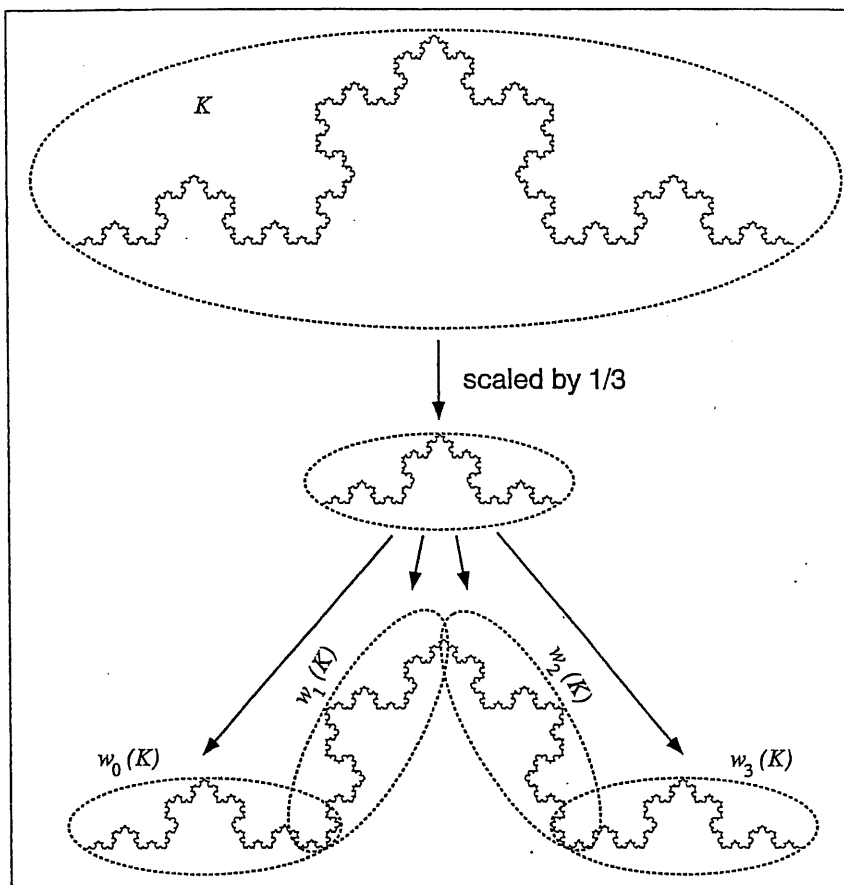


Figure 3.22

The KOCH Collage

The word 'KOCH' is not invariant under W .

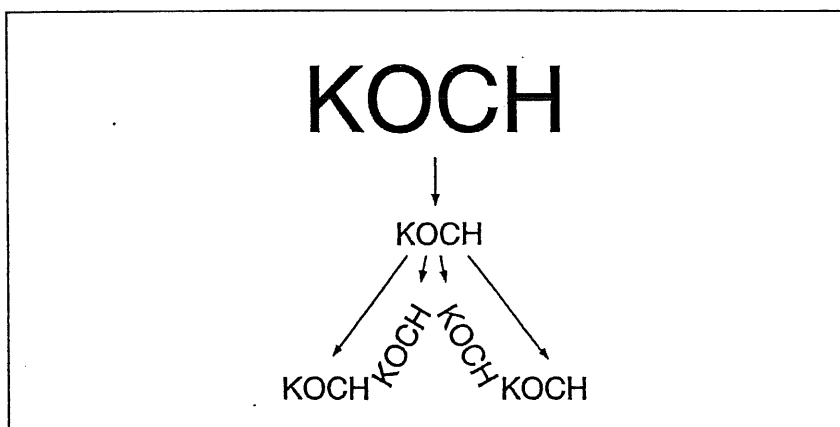
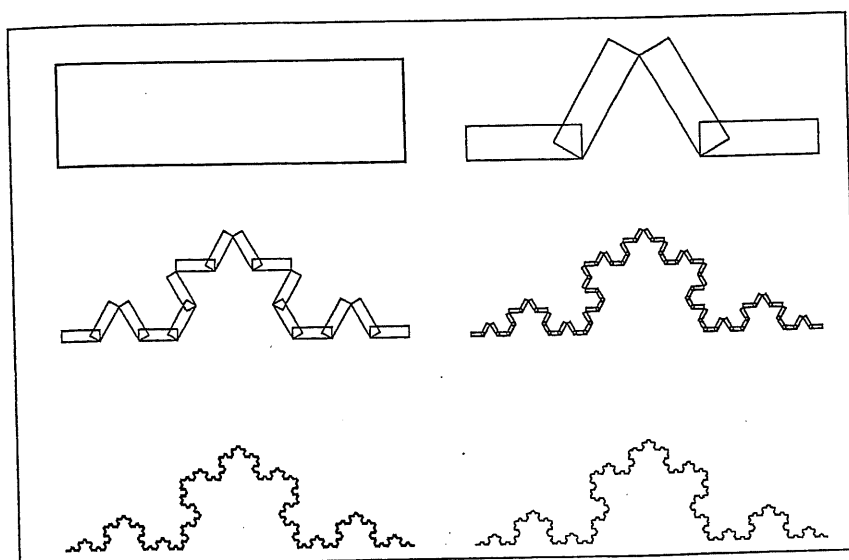


Figure 3.23

then the Koch curve K solves the problem. Moreover, this equation also shows the self-similarity of K since

$$K = w_1(K) \cup w_2(K) \cup w_3(K) \cup w_4(K)$$

states that K is composed of four similar copies of itself. In other words, we have characterized K by its self-similarity. If we further substitute for



Limit Object Koch Curve

Starting with an arbitrary shape, a rectangle, iteration of the Hutchinson operator produces a sequence of images, which converge to the Koch curve.

Figure 3.24

K the collection of the four copies on the right-hand side of the equation, then it becomes clear that K is made of 16 similar copies of itself, and so on. We will come back to this interpretation of self-similarity later in this section.

When we apply the same transformation W to the name KOCH (i.e., X is the image 'KOCH'), we do not get back the name KOCH at all. Rather, we see some strange collage.

Only the Koch Curve Is Invariant Under W

We are led to conjecture that maybe the only image which is left invariant under the collage transformation W is the Koch curve. Indeed, that is a theorem which has far-reaching consequences which will be discussed in chapter 5. A collage transformation like W above is called a *Hutchinson operator*, after J. Hutchinson, who was the first to discuss its properties.²⁴

The Koch Curve As a Limit Object

Having characterized the Koch curve as a fixed point of the Hutchinson operator, we now conclude the analogy to the computation of $\sqrt{2}$ (see eqn. (3.8)). It remains to show that mere iteration of the operator W applied to a starting configuration A_0 yields a sequence

$$A_{n+1} = W(A_n), \quad n = 0, 1, 2, \dots,$$

which converges to the limit object, the Koch curve. This is indeed the case, and figure 3.24 visualizes the limit process, providing pictorial evidence that there is such a self-similar object. Let us summarize.

1. There is a well-defined approximation procedure for the Koch curve, the feedback process

$$A_{n+1} = W(A_n), \quad n = 0, 1, 2, \dots$$

where A_0 can be any initial image and W denotes the Hutchinson operator

$$W(A) = w_1(A) \cup w_2(A) \cup w_3(A) \cup w_4(A)$$

²⁴J. Hutchinson, *Fractals and self-similarity*, Indiana University Math. J. 30 (1981) 713–747.

5.6 Foundation of IFS: The Contraction Mapping Principle

The image coding problem has led us to one of the central questions: how images can be compared or what the distance between two images is. In fact, this is crucial for the understanding of iterated function systems. Without an answer to these questions we will not be able to precisely verify the conditions under which the machine will produce a limiting image. Felix Hausdorff, whom we have already mentioned as the man behind the mathematical foundations of the concept of fractal dimension, proposed a method of determining this distance which is now named after him — the Hausdorff distance. Introducing the Hausdorff distance $h(A, B)$ has two marvelous consequences. First, we can now talk about the sequence of images A_k having the limit A_∞ in a very precise sense: A_∞ is the limit of the sequence A_0, A_1, A_2, \dots provided that the Hausdorff distance $h(A_\infty, A_k)$ goes to 0 as k goes to ∞ . But even more importantly, Hutchinson showed that the operator W , which describes the collage

$$W(A) = w_1(A) \cup w_2(A) \cup \dots \cup w_N(A)$$

is a contraction with respect to the Hausdorff distance. That is, there is a constant c , with $0 \leq c < 1$, such that

$$h(W(A), W(B)) \leq c \cdot h(A, B)$$

for all (compact) sets A and B in the plane. In establishing this fundamental property, Hutchinson was able to inject into consideration one of the most powerful and beautiful principles in mathematics — the contraction mapping principle, which has a long history and owes its final formulation to the great Polish mathematician Stefan Banach (1892–1945).

If the works and achievements of mathematicians could be patented, then the contraction mapping principle would probably be among those with the highest earnings up to now and for the future. Once he allowed himself a certain degree of abstraction, Banach understood that many individual and special cases which floated in the work of earlier mathematicians can be subsumed under one very brilliant principle. The result is nowadays a theorem in *metric topology*, a branch of mathematics which is basic for a great part of modern mathematics and is usually a topic reserved for students of an advanced university-level mathematics courses. We will explain the core of Banach's ideas in a non-rigorous style.

The Hausdorff distance determines the distance of images. It is based on the concept of distance of points to be explained here. Expressed generally, the distance between points of a space X can be measured by a function $d : X \times X \rightarrow \mathbf{R}$. Here \mathbf{R} denotes the real numbers and the function d must have the properties that

- (1) $d(x, y) \geq 0$
- (2) $d(x, y) = 0$ if and only if $x = y$

Measuring Distance: The Metric Space

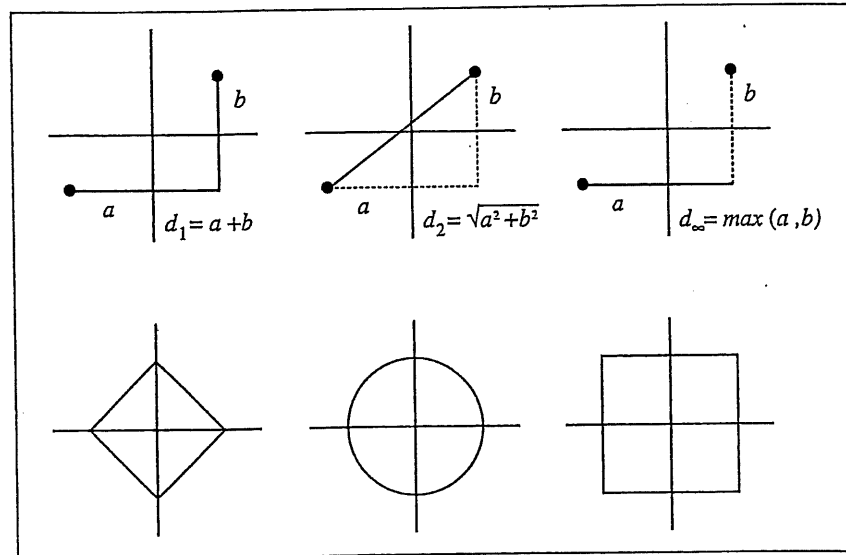


Figure 5.39 : Three methods of measuring distance in the plane (the lattice distance, the Euclidean distance, the maximum norm distance) and the corresponding unit sets (the set of points which have the distance 1 to the origin of the coordinate system).

$$(3) \quad d(x, y) = d(y, x)$$

$$(4) \quad d(x, y) \leq d(x, z) + d(z, y) \text{ (triangle inequality),}$$

hold for all $x, y, z \in X$. We call such a mapping d a *metric*. A space together with a metric is called a *metric space*. Some examples are (see figure 5.39):

(1) For real numbers x and y we can set

$$d(x, y) = |x - y|.$$

(2) For points $P = (x, y)$, $Q = (u, v)$ in the plane we can define

$$d_2(P, Q) = \sqrt{(x - u)^2 + (y - v)^2}.$$

This is the *Euclidean metric*.

(3) Another metric in the plane would be

$$d_\infty(P, Q) = \max\{|x - u|, |y - v|\}.$$

This is the *maximum metric*.

(4) A further metric illustrated in figure 5.39, the *lattice metric*, is given by

$$d_1(P, Q) = |x - u| + |y - v|.$$

The last metric d_1 on the list is sometimes also referred to as the *Manhattan metric*, because it is the distance, a cab driver in Manhattan, New York, would have to drive to get from P to Q .

Once we have a metric for a space X we can talk about *limits* of sequences. Let x_0, x_1, x_2, \dots be a sequence of points from X and a element from X . Then a is the limit of the sequence provided

$$\lim_{k \rightarrow \infty} d(x_k, a) = 0.$$

other words, for any $\varepsilon > 0$ we can find a point x_n in the sequence that any point later in the sequence has distance to a less than ε :

$$d(x_k, a) < \varepsilon, \quad k > n.$$

this case we say that the sequence converges to a . Often it is very desirable to test the convergence of a sequence without knowledge of the limit. This, however, works only if the underlying space X has special nature (i.e., it is a *complete metric space*). Then one may discuss limits by monitoring the distance of consecutive points in the sequence.

The space X is called a *complete metric space* if any Cauchy sequence has a limit which belongs to X . More precisely, this means the following: Let x_0, x_1, x_2, \dots be a given sequence of points in X . It is a Cauchy sequence if for any given number $\varepsilon > 0$ we can find a point x_m in the sequence so that any two points later in the sequence have a distance less than ε :

$$d(x_i, x_j) < \varepsilon, \quad i, j \geq m.$$

When the limit of the sequence exists and is a point of X . Two examples are:

- 1) The set of rational numbers is not complete. There are Cauchy sequences of rational numbers whose limits exist but are not rational numbers. An example of such a sequence is given by

$$x_n = \sum_{k=1}^n \frac{1}{k^2}.$$

This sequence of rational numbers converges to the irrational limit $\pi^2/6$.

- 2) The plane \mathbf{R}^2 is complete with respect to any of the metrics, d_1 , d_2 , or d_∞ .

The Environment of the Contraction Mapping Principle

In chapter 1 we learned that a large variety of dynamic processes and phenomena can be seen from the point of view of a feedback system. A sequence of events a_0, a_1, a_2, \dots is generated starting with an initial event a_0 , which can be chosen from a pool of admissible choices. As time elapses (as n grows), the sequence can show all kinds of behavior. The central problem of dynamical systems theory is to forecast the long-term behavior. Often that behavior will not depend very much on the initial choice a_0 . That is exactly the environment for the contraction mapping principle. It provides everything which we can hope for to make a forecast. But having in mind

the variety of both wild and tame behavior which feedback systems can produce, it is clear that the principle will select some sub-class of feedback systems for which it can be applied. Let us collect the two features which characterize this class:

- (1) **The Space.** The objects — numbers, images, transformations, etc. which we call a_n — must belong to a set in which we can measure the distance between any two of its elements, for example the distance between x and y is $d(x, y)$. Furthermore, the set must be saturated in some sense. That means, if an arbitrary sequence satisfies a special test which examines the possible existence of a limit, then a limit exists and belongs to the set (technically: the space is a *complete metric space*).
- (2) **The Mapping.** The sequence of objects is obtained by a mapping, say f . That means that for any initial object a_0 , a sequence a_0, a_1, a_2, \dots is generated by $a_{n+1} = f(a_n)$, $n = 0, 1, 2, \dots$. Furthermore, f is a contraction. That means that for any two elements of the space, say x and y , the distance between $f(x)$ and $f(y)$ is always strictly less than the distance between x and y .¹¹

For this class of feedback systems the contraction mapping principle gives the following remarkable result:

The Result of the Contraction Mapping Principle

- (1) **The Attractor.** For any initial object a_0 the feedback system $a_{n+1} = f(a_n)$ will always have a predictable long-term behavior. There is an object a_∞ (the limit of the feedback system) to which the system will go. That limit object is the same no matter what the initial object a_0 is. We call a_∞ the unique *attractor* of the feedback system.
- (2) **The Invariance.** The feedback system leaves a_∞ invariant. In other words, if we start with a_∞ , then a_∞ is returned. a_∞ is a fixed point of f , i.e. $f(a_\infty) = a_\infty$.
- (3) **The Estimate.** We can predict how fast the feedback system will arrive close to a_∞ when it is started at a_0 . We only have to test the feedback loop once on the initial object. That means, if we measure the distance between a_0 and $a_1 = f(a_0)$, we can already safely predict how often we have to run the system to arrive near a_∞ within a prescribed accuracy. However, we can estimate the distance between a_0 and a_∞ .

A Contractive

A mapping f is a *contraction* of the metric space X , provided that there is a constant c , $0 \leq c < 1$, such that for all x, y in X one has that

$$d(f(x), f(y)) \leq cd(x, y).$$

The constant c is called the *contraction factor* for f . Let a_0, a_1, a_2, \dots be a sequence of elements from a complete metric space X defined by $a_{n+1} = f(a_n)$. The following holds true:

$d(a_n, a_m)$ with a constant $0 \leq c < 1$.

The ε -Collar

The ε -collar of a set A in the plane. Note that the ε -collar of A includes A and is not just a set of points close to A , as the term 'collar' might suggest.

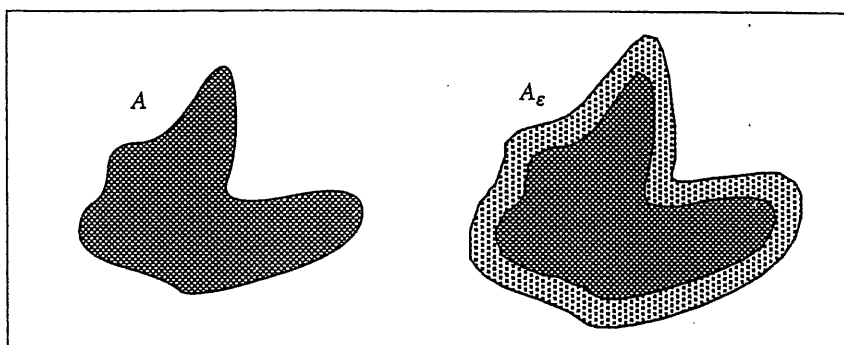


Figure 5.40

Definition of the Hausdorff Distance

In precise mathematical terms the definition of the Hausdorff distance is as follows. Let X be a complete metric space with metric d . For any compact subset A of X and $\varepsilon > 0$, define the ε -collar of A by

$$A_\varepsilon = \{x \in X \mid d(x, y) \leq \varepsilon \text{ for some } y \in A\}.$$

For any two compact subsets A and B of X the Hausdorff distance is

$$h(A, B) = \inf\{\varepsilon \mid A \subset B_\varepsilon \text{ and } B \subset A_\varepsilon\}.$$

According to Hausdorff the space of all compact subsets of X , equipped with the Hausdorff distance, is another complete metric space. This implies that the space of all compact subsets of X is a suitable environment for the contraction mapping principle.

With this definition it follows that $h(A, B) = 0$ when A is equal to B . Also, if A is just a point and B is just a point, then $h(A, B)$ is the distance between A and B in the ordinary sense. Figure 5.41 illustrates that fact and gives a few more examples useful for getting acquainted with the notion of Hausdorff distance.

Let us now return to the state of affairs which Hutchinson obtained when analyzing the operator W

$$W(A) = w_1(A) \cup w_2(A) \cup \dots \cup w_N(A),$$

where the transformations $w_i, i = 1, \dots, N$, are contractions with contraction factors c_i . Hutchinson was able to show that W is also a contraction, however, with respect to the Hausdorff distance. Thus, the contraction mapping principle can be applied to the iteration of the Hutchinson operator W . Consequently, whatever initial image is chosen to start the iteration of the IFS, for example A_0 , the generated sequence

$$A_{k+1} = W(A_k), \quad k = 0, 1, 2, 3, \dots$$

The Hutchinson Operator

Four Examples of Hausdorff Distance

To obtain the Hausdorff distance between two planar sets A and B we compute $a_\varepsilon = \inf\{\varepsilon \mid B \subset A_\varepsilon\}$ (left figures) and $b_\varepsilon = \inf\{\varepsilon \mid A \subset B_\varepsilon\}$ (right figures). B barely fits into the a_ε -collar of A , and A barely fits into the b_ε -collar of B . The Hausdorff distance is the maximum of both values, $h(A, B) = \max\{a_\varepsilon, b_\varepsilon\}$. The sets A and B are two points (top row), a disk and a line segment (second row), a disk and a large square (third row, here $b_\varepsilon = 0$), and two intersecting disks (bottom row).

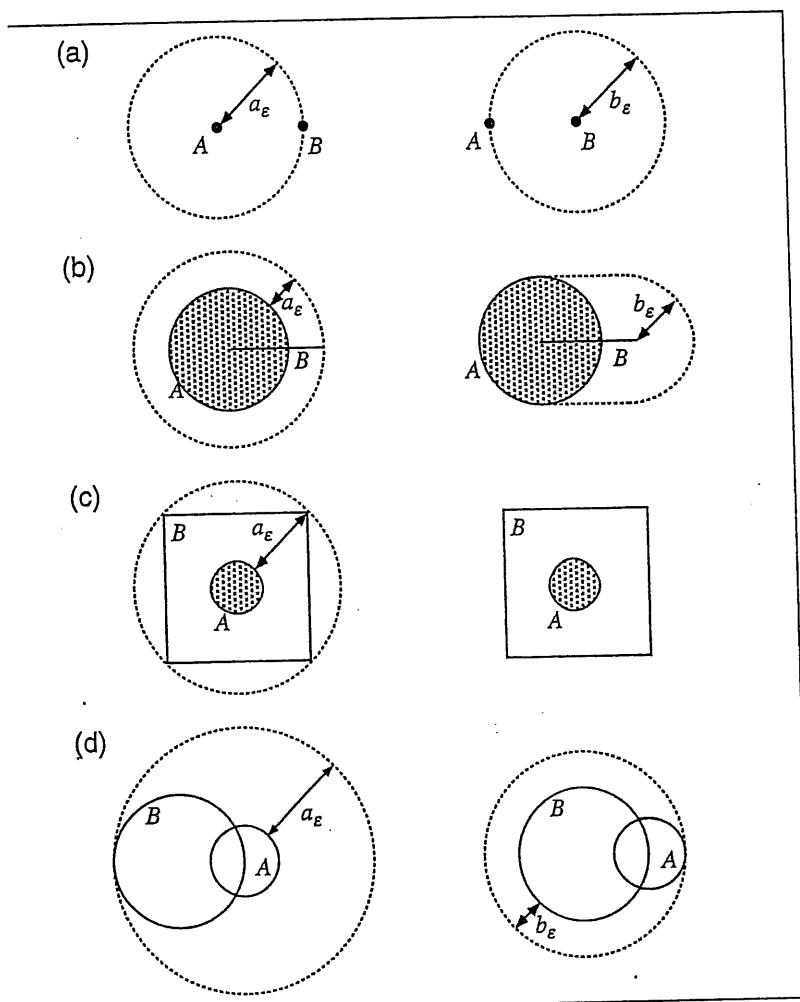


Figure 5.41

will tend towards a distinguished image, the attractor A_∞ of the IFS. Moreover, this image is invariant:

$$W(A_\infty) = A_\infty.$$

This solves a central problem raised in chapter 3. The Koch curve, the Sierpinski gasket, etc. all seem to be objects in the plane, and there are convergent processes for them, namely the iteration of the corresponding Hutchinson operators. But we could not prove that these fractals really exist and are not just some impossible artifact of a self-referential scheme such as the assumption of a barber who shaves all men who do not shave themselves — obviously a falsehood. However, now, with Hutchinson and Hausdorff's results in hand, we can be sure that the limit object with the extraordinary self-similarity property truly exists.

The contraction mapping principle even gives us something in addition for free. Knowing the contraction factor c of the Hutchinson operator W , we can estimate how fast the IFS will produce the final image from just applying

