Chapter 1

Fractals

I.1 Self-Similarity

Definition 1.1 *A set* $\Omega \subset \mathbb{R}^n$ *is* self-similar (or affine self-similar) *if there is a proper subset* $\tilde{\Omega} \subset \Omega$*, a linear transformation* T*, and a vector* $v \in \mathbb{R}^n$ *such that* $T(\tilde{\Omega}) + v = \Omega$.

Recall that for any set $\Omega \subset \mathbb{R}^n$ and any linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$, $T(\Omega) = \{T(x) \,|\, x \in \Omega\}$, and for any vector $v \in \mathbb{R}^m$, $T(\Omega) + v = \{T(x) + v \,|\, x \in \Omega\}$; see Figure 1.We call the vector v the *shift vector* figure (it shifts the set $T(\Omega)$).

That is, Ω is self-similar if it is made up of smaller, linearly distorted, copies of itself. Note that a self-similar set is either trivially complicated or infinitely complicated. An example of the former is a square (eg., it is composed of 4 equal sub-squares). Examples of the latter ('fractals') will be given in the next section.

More generally, if $\mathcal{J}$ is a set of functions we say that Ω is $\mathcal{J}$ *self-similar* if there is a proper subset $\tilde{\Omega} \subset \Omega$ and an $f \in \mathcal{J}$ such that $f(\tilde{\Omega}) = \Omega$. This more general notion of self-similarity arises for example in the study of the Mandelbrot set and Julia sets.

I.2 Some fractals

I.2.1 The Cantor Set

Construction of the Cantor Set

Let $\mathcal{C}_0 = [0,1]$, $\mathcal{C}_1 = [0,\frac{1}{3}] \cup [\frac{2}{3},1]$, $\mathcal{C}_2 = [0,\frac{1}{9}] \cup [\frac{2}{9},\frac{1}{3}] \cup [\frac{2}{3},\frac{7}{9}] \cup [\frac{7}{9},1]$, and in general, let $\mathcal{C}_{n+1}$ be the union of the 2^{n+1} closed intervals, each of length $(\frac{1}{3})^{n+1}$, obtained by removing the open middle thirds of the 2^n closed intervals of $\mathcal{C}_n$. We define the Cantor set $\mathcal{C}$ to be the intersection of all the $\mathcal{C}_n$; $\mathcal{C} = \cap_{n=0}^{\infty} \mathcal{C}_n$. Another way to describe this is to say that $\mathcal{C}$ is the set of points in $[0,1]$ that remain after removing the open middle third interval $(\frac{1}{3},\frac{2}{3})$, and then removing the open middle thirds from the remaining two closed intervals $[0,\frac{1}{3}]$, $[\frac{2}{3},1]$, and then removing the open middle thirds from the remaining four closed intervals, etc ad infinitum. Note that the sets $\mathcal{C}_n$ are approximations of $\mathcal{C}$ in the sense that $\mathcal{C} = \lim_{n\to\infty} \mathcal{C}_n$, so we can get an impression of what $\mathcal{C}$ looks like by looking at $\mathcal{C}_n$ as n gets large (see Figure 2). figure

This construction of $\mathcal{C}$ removes infinitely many intervals from $[0,1]$, so we might wonder if there are any points in $\mathcal{C}$. Some obvious points are the end points of the open middle third intervals that were removed; $\{0, 1, \frac{1}{3}, \frac{2}{3}, \frac{1}{9}, \frac{2}{9}, \frac{7}{9}, \frac{8}{9}, \ \dots\}$. So there are at least a countably infinite number of points in $\mathcal{C}$. In fact, we will see that there are *many* more points in $\mathcal{C}$ than these. But let's first describe some of the properties of the Cantor set.

Since $[0,1] = \mathcal{C} \cup \{\text{intervals removed}\}$ (a disjoint union, notice), the length of $\mathcal{C} = 1 -$ (total length of the intervals removed). The length removed in the first stage is $\frac{1}{3}$, the length removed in the second stage is $2 \cdot (\frac{1}{3})^2$, the length removed in the third stage is $2^2 \cdot (\frac{1}{3})^3$, etc, so the total length of the intervals removed is $\sum_{n=1}^{\infty} 2^{n-1}(\frac{1}{3})^n = \frac{1}{2}\sum_{n=1}^{\infty}(\frac{2}{3})^n = \frac{1}{2}\left(\frac{2/3}{1/3}\right) = \frac{1}{2}(2) = 1$. Thus, the length of $\mathcal{C}$ is 0. This implies that $\mathcal{C}$ cannot contain any intervals, i.e., that is is 'dust' (between any two points in $\mathcal{C}$ is a point that is not in $\mathcal{C}$).

We've noted that the end points of the intervals removed during the construction are in $\mathcal{C}$, but what other points are in $\mathcal{C}$, if any? It's very difficult to see what other points are in $\mathcal{C}$ by relying on this geometric construction. For example, can you see why the numbers $\frac{1}{4}$ and $\frac{3}{4}$ are in $\mathcal{C}$? They are not the end points of any interval that was removed, yet they are never removed in the construction of $\mathcal{C}$. To see exactly what numbers are in $\mathcal{C}$, it's much more convenient to represent numbers in a way that reflects the structure of $\mathcal{C}$. Going back to the construction, note that the points in $\mathcal{C}_1$ are precisely the numbers in $[0,1]$ that have no 1 in the first place of their ternary expansion (here we need to resolve the ambiguity of $[\frac{1}{3}]_3$ and $[\frac{2}{3}]_3$; we choose $[\frac{1}{3}]_3 = 0.0\overline{22}$ and $[\frac{2}{3}]_3 = 0.2$, and similarly for the other end points; we choose the representation that contains only $2's$). Similarly, $\mathcal{C}_2$ are the numbers in $[0,1]$ that have no 1 in either the first or second places of their ternary expansion. So we see that $\mathcal{C}_n$ is precisely the numbers in $[0,1]$ that have no 1 in any of the first n places of their ternary expansion. Thus, since $\mathcal{C} = \lim_{n\to\infty} \mathcal{C}_n$, the numbers in $\mathcal{C}$ are the numbers in $[0,1]$ that have no 1 in their ternary expansion; see Figure 3. figure

For example, $[\frac{1}{4}]_3 = 0.02\overline{02}$ and $[\frac{3}{4}]_3 = 0.2\overline{02}$ so both $\frac{1}{4}$ and $\frac{3}{4}$ are in $\mathcal{C}$. Moreover, if $a = b_1 b_2 b_3 \dots$ is

a sequence of $0's$ and $2's$, then the number x with $[x]_3 = 0.a$ is a number in $\mathcal{C}$. Just how many of these numbers are there? To answer this we observe that we can match elements in the set $\mathcal{B}$ of all sequences of the form $0.b$ where b is a sequence of $0's$ and $1's$, with numbers in $[0,1]$ via binary expansions (however, not in a one-to-one manner). That is, if $0.b$ is any sequence of $0's$ and $1's$, then there is a number $x \in [0,1]$ such that $[x]_2 = 0.b$, in fact $x = \frac{b_1}{2} + \frac{b_2}{2^2} + \frac{b_3}{2^3} + \cdots$. Now, the set $\mathcal{S}$ of all sequences of the form $0.a$ where a is a sequence of $0's$ or $2's$ has the same cardinality as the set $\mathcal{B}$; just match each $b \in \mathcal{B}$ with an element $a \in \mathcal{S}$ by changing every 1 in b to a 2, and visa versa, match each element $a \in \mathcal{S}$ with an element $b \in \mathcal{B}$ by changing every 2 in a to a 1. Since $\mathcal{B}$ has the same cardinality as $[0,1]$, and since $\mathcal{S}$ has the same cardinality as $\mathcal{C}$, the Cantor set $\mathcal{C}$ has the same cardinality as the interval $[0,1]$! This seems bizarre because in some sense $\mathcal{C}$ is a 'small' subset of $[0,1]$ (it is a subset of length 0). This shows you that by 'rearranging' the points in $[0,1]$ we can obtain a set of length zero (there are also 'generalized Cantor sets' which have lengths anywhere between 0 and 1, so more generally we can rearrange the points in $[0,1]$ to obtain a set of any length between 0 and 1, including 0 and 1; cf. Excercise xx). exercise

How are the numbers in $[0,1]$ rearranged to obtain $\mathcal{C}$? The discussion in the previous paragraph explained how we could determine the cardinality of $\mathcal{C}$ by matching each number in the interval $[0,1]$ with a number in $\mathcal{C}$ in a one-to-one manner;

$$[0,1] \ni x \mapsto b = [x]_2 \mapsto a \in \mathcal{S} \mapsto y = \frac{a_1}{3} + \frac{a_2}{3^2} + \cdots \in \mathcal{C} \tag{2.1}$$

(we make the convention that if x has two binary expansions, we take the one that ends in zeros). If you look more closely at this matching, you'll see that some numbers in $\mathcal{C}$ are actually missed. For example, $\frac{1}{3} \in \mathcal{C}$ is not matched with any number in $[0,1]$; $[\frac{1}{3}]_3 = 0.02\overline{22}$, but $\frac{1}{2} \in [0,1]$ is mapped to $\frac{2}{3} \in \mathcal{C}$ (see exercise xxx). So this matching actually only uses a strict subset of $\mathcal{C}$ (which is sufficient to prove that the exercise cardinality of $\mathcal{C}$ is at least as large as the cardinality of $[0,1]$). However, 'most' of the numbers of $\mathcal{C}$ are matched with a number in $[0,1]$ (exercise xx), so this way to match the two sets gives us a good impression exercise of how a rearrangement of $[0,1]$ can produce $\mathcal{C}$.

Generally, we can represent any rearrangement of $[0,1]$ by drawing the graph of the function $\varphi(x)$ that represents the rearrangement (i.e., $\varphi(x) = y$ means the rearrangement moves x to y). Now, it's no mystery how one can rearrange $[0,1]$ to obtain a set of small length. Let ε be any small positive number. Then the function $\varphi_\varepsilon(x) = \varepsilon x$ rearranges $[0,1]$ into a set of length ε, namely the set $[0,\varepsilon]$. Notice that the slope of the graph of $\varphi_\varepsilon(x)$ is small; the slope of the graph of any function that rearranges $[0,1]$ into a set of small length must necessarily be rather small. Since the length of $\mathcal{C}$ is zero, the graph of the function $\varphi(x)$ that represents the rearrangement of $[0,1]$ into $\mathcal{C}$ must in some sense have zero slope. But the graph of this function cannot be flat on any interval because we know that if $x_1 \neq x_2$, then $\varphi(x_1) \neq \varphi(x_2)$. So it's not obvious what the graph of this function looks like; it begins at $(0,0)$, ends at $(1,1)$, its range is $\mathcal{C}$ (so if you projected the graph onto the y-axis it would be $\mathcal{C}$), and is 'flat'!

To get an idea of what that graph looks like, let's define a function $\varphi_{\mathcal{C}}(x)$ which matches the numbers in $[0,1]$ to numbers in $\mathcal{C}$ as described above in equation (*). Since 'most' numbers in $\mathcal{C}$ are matched in this way with a number in $[0,1]$, the graph of $\varphi_{\mathcal{C}}(x)$ will give an accurate impression of the way $[0,1]$ is rearranged to make $\mathcal{C}$.

Figure 4 shows the graph of $\varphi_{\mathcal{C}}(x)$. It was obtained by taking $x \in [0,1]$, computing $[x]_2$, changing every 1 in $[x]_2$ to a 2, then summing up the resulting ternary expansion to obtain $y = \varphi_{\mathcal{C}}(x)$. If you look closely you'll see that the graph appears to be flat everywhere, but also has lots of jumps. The jumps are precisely at the points x where $x = (\frac{m}{2^n})$ for some positive integer n, and positive integer $m < 2^n$ (these are the numbers which, by the convention mentioned above, have binary expansions that end in zeros). Note that these points are dense in $[0,1]$, so the graph of $\varphi_{\mathcal{C}}(x)$ has a jump almost everywhere, and is 'flat' everywhere else. figure

Let $\mathcal{E} = \{0, 1, \frac{1}{3}, \frac{2}{3}, \frac{1}{9}, \frac{2}{9}, \frac{7}{9}, \frac{8}{9}, \ \dots \}$ be the set of edges of the intervals removed in the construction of the Cantor set $\mathcal{C}$.

Claim: $\mathcal{E}$ is a 'small' subset of $\mathcal{C}$, i.e., 'most' of the numbers in $\mathcal{C}$ are not edge points.

Proof: If $x \in \mathcal{E}$, then $[x]_3$ ends in $\overline{00}$ because $x = \frac{m}{3^n}$ for some positive integer $m < 3^n$ (exercise xx).If $\mathcal{S}_o$ is the subset of $\mathcal{S}$ of sequences that end in $\overline{00}$, then $\mathcal{S}_0$ is a 'small' subset of $\mathcal{S}$ in the sense that the cardinality of $\mathcal{S} \setminus \mathcal{S}_0$ is the same as the cardinality of $\mathcal{S}$ (exercise xx) $\square$ exercise exercise

So we could have removed the open *closed* middle thirds in the construction of $\mathcal{C}$ and still have obtained essentially $\mathcal{C}$. However, the set $\mathcal{E}$ shows us where the points of $\mathcal{C}$ are.

Claim: $\overline{\mathcal{E}} = \mathcal{C}$.

Proof: Exercise xx. exercise

In other words, the edge points $\mathcal{E}$ accumulate to $\mathcal{C}$; if $x \in \mathcal{C}$ is any point in the Cantor set, then there is an infinite sequence of points from $\mathcal{E}$ that converge to x. So although $\mathcal{E}$ is a negligibly small subset of $\mathcal{C}$, the edge points do show us exactly where the points in $\mathcal{C}$ are, and so sketching the edge points gives us an accurate impression of what $\mathcal{C}$ looks like (however, sketching $\mathcal{E}$ is no easy task!).

Now we turn to the self-similarity of the Cantor set. We define the interval $A_{a_1 a_2 \cdots a_n}$ to be the interval that contains all numbers in $[0,1]$ whose ternary expansion begins with $a_1 a_2 \cdots a_n$;

$$\begin{aligned} A_{a_1 a_2 \cdots a_n} &= \{x \in [0,1] \mid [x]_3 = 0.a_1 a_2 \cdots a_n a_{n+1} a_{n+2} \cdots \} \\ &= \{x \in [0,1] \mid [x]_3 = 0.a_1 a_2 \cdots a_n \, \vec{a} \, \} \end{aligned}$$

where $\vec{a}$ is an arbitrary (infinitely long) sequence of $0's$, $1's$, and $2's$. For example,

$$\begin{aligned} A_0 &= [0,\, 1/3] \\ A_1 &= [1/3,\, 2/3] \\ A_2 &= [2/3,\, 1] \\ A_{00} &= [0,\, 1/9] \\ A_{22} &= [8/9,\, 1] \\ A_{022} &= [8/27,\, 9/27] \end{aligned}$$

Then we define the sets $\mathcal{C}_{a_1a_2\cdots a_n}$ to be those parts of the Cantor set that lie in $A_{a_1a_2\cdots a_n}$;

$$\begin{aligned} \mathcal{C}_{a_1a_2\cdots a_n} &= \mathcal{C} \cap A_{a_1a_2\cdots a_n} \\ &= \{x \in \mathcal{C} \mid [x]_3 = 0.a_1a_2\cdots a_n\,\vec{c}\,\} \end{aligned}$$

where $\vec{c}$ is an arbitrary sequence of $0's$ and $2's$.

Let $3^m\,\mathcal{C}_{a_1a_2\cdots a_n} = \{x \,|\, x = 3^m y \text{ for some } y \in \mathcal{C}_{a_1a_2\cdots a_n}\}$. That is, $3^m\,\mathcal{C}_{a_1a_2\cdots a_n}$ are the numbers in $\mathcal{C}_{a_1a_2\cdots a_n}$ multiplied by 3^m. Now recall that $[3^m x]_3$ is the ternary expansion of x shifted to the left by m places for positive m (see Exercise xx in Appendix 1). Therefore, $3^n\,\mathcal{C}_{a_1a_2\cdots a_n} = \{x \mid [x]_3 = a_1a_2\cdots a_n\,.\,\vec{c}\,\}$. exercise Now if we let $z = a_1 3^{n-1} + a_2 3^{n-2} + \cdots + a_{n-1}3 + a_n$ (so that $[z]_3 = a_1a_2\cdots a_n$), then

$$3^n\,\mathcal{C}_{a_1a_2\cdots a_n} - z = \{x \mid [x]_3 = 0.\vec{c}\,\}$$

where $\vec{c}$ is an arbitrary sequence of $0's$ and $2's$ (this is because if x is a number such that $[x]_3 = 0.a_1a_2\cdots a_n\,\vec{c}$, then $[3^n x - z]_3 = 0.\vec{c}\,$).

In other words, the set of numbers $3^n\,\mathcal{C}_{a_1a_2\cdots a_n} - z$ are precisely those numbers in $[0,1]$ whose ternary expansion contains only $0's$ and $2's$. That is, $3^n\,\mathcal{C}_{a_1a_2\cdots a_n} - z$ is the Cantor set. This demonstrates the self-similarity of the Cantor set (it is similar to pieces of itself after magnifying and shifting the pieces). For example,

$$\begin{aligned} 3\,\mathcal{C}_2 - 2 &= \mathcal{C} \\ 9\,\mathcal{C}_{22} - 8 &= \mathcal{C} \\ 27\,\mathcal{C}_{022} - 8 &= \mathcal{C} \end{aligned}$$

This shows that $\mathcal{C}$ is self-similar, and identifies the self-similar pieces of $\mathcal{C}$ (here, the linear transformation T and vector z mentioned in Definition 1.1 are 3^n and z respectively).

Summary of properties of the Cantor set:

- the length of $\mathcal{C}$ is zero
- $\mathcal{C}$ is totally disconnected (is 'dust')
- $\mathcal{C}$ is a closed set
- $\mathcal{C}$ has the same cardinality as $[0, 1]$
- every point in $\mathcal{C}$ is a limit of a sequence of end points $\mathcal{E}$
- $\mathcal{C}$ is self-similar

I.2.2 The Sierpinski Triangle

Construction

Here we construct a fractal in $\mathbb{R}^2$. The construction is similar to the construction of the Cantor set; we successively remove parts of an initial set. Here we start with a solid triangle T_0. We divide that triangle into 4 equal equilateral triangles of $\frac{1}{2}$ the size (and so $\frac{1}{4}$ the area) of the original solid triangle and remove the middle triangle; T_1. Now we do the same for each of the remaining 3 solid triangles; divide them up into 4 equal triangles and remove the middle triangle, etc (see Figure 5); T_2. The Sierpinski triangle T is the limit of these sets; $T = \lim_{n\to\infty} T_n$. Or, since $T_{n+1} \subset T_n$, it is the intersection of all these sets; $T = \cap_0^\infty T_n$. figure

Now let's see what's left after carrying out this procedure ad infinitum. First of all one can show that the total area removed is equal to the area A_o of the initial triangle. But just as with the Cantor set, this doesn't mean there isn't anything left at the end. Since we removed open triangles, all the edges (lines) of the triangles removed are left. These edges form a curve that is infinitely long (exercise xx). exercise

To quantify the self-similarity of the Sierpinski triangle T, let $A_{a_1a_2\ldots a_n}$ denote the region of T_o with address $a_1a_2\ldots a_n$ and $T_{a_1a_2\ldots a_n}$ denote the part of T inside $A_{a_1a_2\ldots a_n}$ (see Figure 14). Then we see that $2^n\,\mathbb{I}\,T_{a_1a_2\ldots a_n} + v = T$, for appropriate shift vector v, and where $\mathbb{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. figure

I.2.3 More Fractals

The von Koch curve

Unlike the Cantor set and Sierpinski's triangle, the von Koch curve is not constructed by successively removing pieces from an initial set, but by adding to (and modifying) an initial curve. We begin with a line segment of length 1 (say); $K_0 = [0, 1]$. Then we remove the open middle third segment $(\frac{1}{3}, \frac{2}{3})$ and add two line segments of length $\frac{1}{3}$ forming an equilateral triangle. This leaves us with 4 lines segments K_1 of length $\frac{1}{3}$; Figure 6. We continue with this procedure; for each line segment we remove the open middle third figure

segment and form an equilateral triangle. The von Koch curve K is the limiting curve; $K = \lim_{n\to\infty} K_n$[1].

The length of an intermediate curve K_n is $4^n(\frac{1}{3})^n$, and so the length of K is $\lim_{n\to\infty} 4^n(\frac{1}{3})^n = \infty$; the von Koch curve is infinitely long. The whole curve resides in a region of finite area, so to fit it in it would have to be very complicated! To get an idea of how complicated the curve really is, note that in constructing the von Koch curve, we successively removed the open middle third intervals of the line segment K_0. This is exactly what we did to construct the Cantor set, so $K \cap K_0 = \mathcal{C}$. Furthermore, at each end point of the intervals removed there is a corner. We saw above (§1.2.1) that the end points $\mathcal{E}$ are dense in $\mathcal{C}$. Since each 'side' of the von Koch curve is self-similar to $K \cap K_0$, there is a corner almost everywhere along K. One can show that K is a continuous curve (Exercise xx), but since there is a corner almost everywhere along the curve, K is non-differentiable everywhere. exercise

Instead of adding a triangle to the intervals removed, we could add a square. This results in a 'square' von Koch curve; see Figure 7. figure

[1]The sense in which the curves K_n converge to K can be made precise using the Hausdorff metric h described below; $\lim_{n\to\infty} h(K_{n+1}, K_n) < (\frac{1}{3})^n$, so $\{K_n\}$ is a Cauchy sequence in $(\mathcal{X}_2, h)$, the metric space of images in $\mathbb{R}^2$, and so has a limit, which we call K.