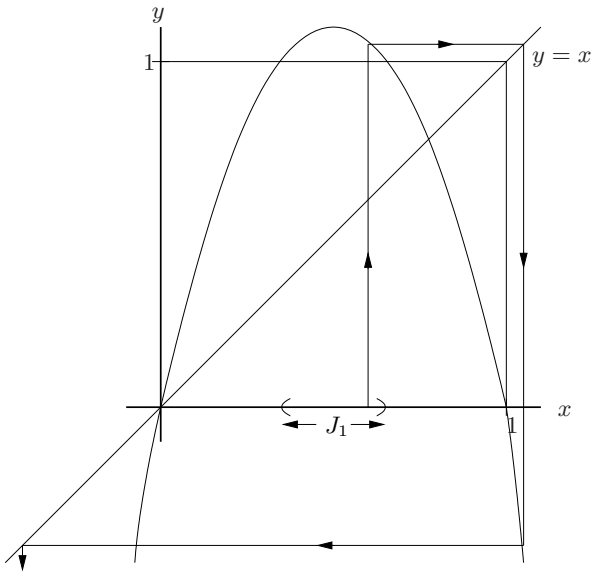
Lecture 24

A return to the dynamics of the Logistic Map: The case $a > 4$

We now examine the interesting iteration dynamics associated with the logistic map when $a > 4$. The graph of $f_a(x) = ax(1-x)$ for $a \cong 4.5$ is sketched below:



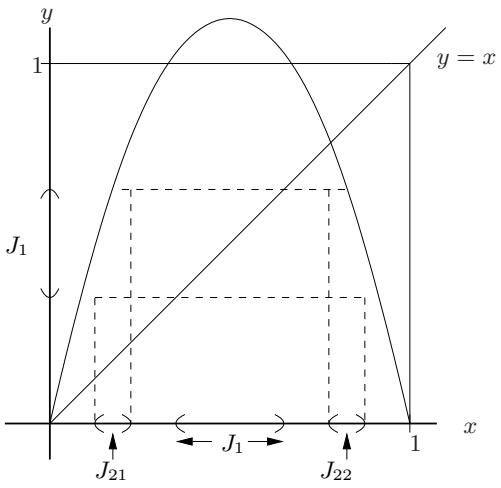
The most important qualitative feature of this graph is that f_a no longer maps $[0,1]$ into itself: There is an open interval of points, J_1, that are mapped outside the interval $[0,1]$. For $x \in J_1$, $f(x) > 1$. Note what happens to points $x \in J_1$ under further iteration of f, as shown in the figure. We note from the graph that: 1) $f^2(x) < 0$, 2) $f^{n+1}(x)$ for $n \geq 2$ and finally 3) $f^n(x) \to -\infty$ as $n \to \infty$.

The natural question that follows is "Are there any other points in $[0,1]$ that eventually leave $[0,1]$?" The answer is "Yes". Note that every point $x \in [0,1]$ has two preimages $y_1 \neq y_2$ such that $f(y_1) = f(y_2) = x$. This leads us to consider the set J_2 of all preimages of the set J_1, i.e.

$$J_2 = \{x \in [0,1] \mid f(x) \in J_1\} \tag{1}$$

Since $x \in J_2$ implies that $f(x) \in J_1$, this implies in turn that $f^2(x) \notin [0,1]$ and that $f^n(x) \to -\infty$ as $n \to \infty$. Now, what is the set J_2, the set of preimages of J_1?

We may determine the set J_2 graphically by "copying" the set J_1 onto the y-axis and then applying f "backwards", i.e. using the line $y = x$ to determine which points on the x-axis are mapped to J_1. This is shown below:



Thus, $f(J_{21}) = J_1$ and $f(J_{22}) = J_1$. The set $J_2 = J_{21} \cup J_{22}$ is the set of all points $x \in [0,1]$ such that $f^2(x) \notin [0,1]$, i.e. the set of points that leave $[0,1]$ after two applications of f.

The reader should see the pattern now. We define the set J_3 of all preimages of J_2, i.e.

$$J_3 = \{x \in [0,1] \mid f(x) \in J_2\}\,. \tag{2}$$

Note that $x \in J_3$ implies that $f^2(x) \in J_1$ which in turn implies that $f^3 \notin [0,1]$. Now continue this process, if possible, to define the following sets of points,

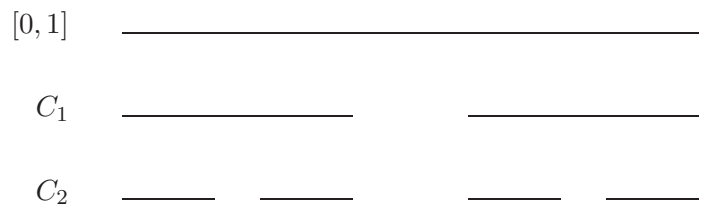
$$J_n = \{x \in [0,1] \mid f^n(x) \notin [0,1]\}\,,\ n \geq 1. \tag{3}$$

It is now convenient to define the following sets:

$$\begin{aligned} C_1 &= [0,1] - J_1 && (4)\\ C_2 &= C_1 - J_2 = [0,1] - (J_1 \cup J_2) && (5)\\ &\vdots \\ C_n &= C_{n-1} - J_n = [0,1] - \left(\bigcup_{k=1}^{n} J_k\right). && (6) \end{aligned}$$

From the definition of the J_k, we see that for $n \geq 1$, C_n is the subset of points in $[0,1]$ that remain in $[0,1]$ after n iterations of f. The natural question is, "Is there a set of points J that remain on $[0,1]$ after any number of iterations of f?" In order to answer this, let us examine the sets C_1, C_2, etc.

From the figures shown earlier, and the definitions in (6), the sets C_1 and C_2 have the following structure:



Note that C_1 is obtained by means of a "dissection procedure" – a removal of the open set J_1 from $[0,1]$. C_2 is obtained from a "dissection" of C_1 – a removal of an open set from each of the subintervals making up C_1. Note that $[0,1]$, C_1 and C_2 are **closed intervals**. The construction of intervals C_1 and C_2 by means of a "dissection" procedure is reminiscent of the "middle-thirds dissection" procedure that was used to construct the ternary Cantor set discussed earlier in the course.
At this time, we simply state, without proof, the following result:

> The sequence of sets $C_1, C_2, \ldots$ converges, in the limit $n \to \infty$, to a "Cantor-like set" $C \subset [0,1]$, i.e. $\lim_{n\to\infty} C_n = C$. For any $x \in C$, $f^n(x) \in J$ for all $n \geq 0$. (The term "Cantor-like" will be defined shortly.)

In other words, **the set of points C that remain in $[0,1]$ after any number of iterations of f_a is a "Cantor-like set"**. Note that the structure of this set, i.e. the positions of points $x \in C$, $x \notin \{0,1\}$, is dependent upon the logistic map parameter a. For example, the size/length of the open interval J_1 removed form $[0,1]$ to produce J_1 is dependent upon a: As $a \to 4^+$, this interval is smaller. Of course, at $a = 4$, no dissection takes place. The reader is encouraged to find endpoints of the intervals that make up the set J_1, as functions of the parameter a, hence the length of the removed interval. This size will also determine the sizes of the sets J_{21} and J_{22} removed in the next dissection procedure, although not in a linear way since the map $f_a(x)$ is not a linear function.

The reader is invited to consider an alternate definition of the set C constructed above:

$$C = \{x \in [0,1] \mid x \text{ is a periodic point of } f_a \text{ with period } n,\ n \geq 1\}\,.$$

Graphical analysis should show (as indeed our previous analysis of the bifurcations of f_a did) that all periodic orbits of f_a are repulsive, as they were for the case $a = 4$.

Cantor (or "Cantor-like") sets

We shall be studying Cantor-like sets in a little more detail later in the course. But while we are on the subject, let us once again take a look at the famous "ternary Cantor set" because of its historical importance in mathematics. Another reason is because it is a little simpler in structure since the dissection procedure is linear, as opposed to the nonlinear procedure associated with the logistic maps examined earlier.

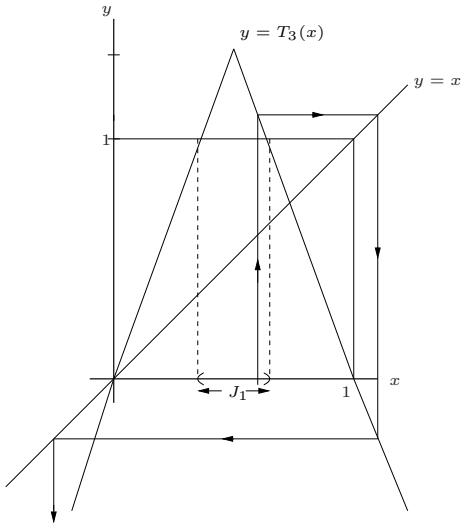
We don't really have to study Cantor-like sets looking at the iteration of functions, since Cantor-like sets may be produced by a limiting procedure of dissection, i.e., removal of points from intervals/sets. But it is instructive to consider iteration since, after all, we have been examining the iteration of functions for the last few weeks.

We shall consider a family of linearized versions of the logistic maps $f_a(x)$, namely, the following family of modified Tent maps,

$$T_a(x) = \begin{cases} ax & 0 \le x \le \frac{1}{2} \\ a(1-x) & \frac{1}{2} < x \le 1\,. \end{cases} \tag{7}$$

The maximum value of $T_a(x)$ is $T_a\left(\frac{1}{2}\right) = \frac{a}{2}$. When $a = 2$, $T_2(x)$ is the Tent Map that we have studied in past lectures.

In the case $a = 3$, the graph of the Tent Map extends out of the "box", i.e., $[0,1] \times [0,1]$, as shown below. We see that T_3 maps some points in [0,1] out of [0,1]. The fate of these points, as in the case of the logistic map f_a for $a > 4$ is that under further iteration, they go to $-\infty$.



We'll proceed as we did for the logistic map f_a in the previous section by making a few observations:

1. Points in the (open) interval
$$J_1 = \left(\frac{1}{3}, \frac{2}{3}\right) \tag{8}$$
are mapped out of $[0,1]$ by T_3. In other words, these points leave $[0,1]$ after one application of T_3.

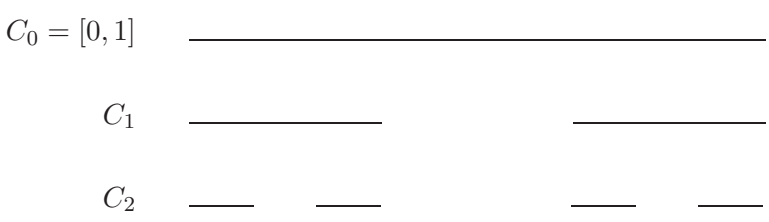
2. Points in $[0,1]$ that are mapped by T_2 to the interval $\left(\frac{1}{3}, \frac{2}{3}\right)$ are mapped out of $[0,1]$ after one additional application of T_3. These points lie in the **preimage** of the interval $\left(\frac{1}{3}, \frac{2}{3}\right)$ which, as in the case of the logistic map, can be determined graphically by backwards iteration to be the set
$$J_2 = \left(\frac{1}{9}, \frac{2}{9}\right) \bigcup \left(\frac{7}{9}, \frac{8}{9}\right) . \tag{9}$$
In other words, points in J_2 leave $[0,1]$ after two applications of T_3.

This procedure can be continued, but it is perhaps easier to focus on the points in [0,1] which remain in [0,1] after a given number of iterations. As we did in the previous lecture for the logistic maps, we determine the sets of points in $[0,1]$ which remain in $[0,1]$ after n applications of the map T_3:

1. The set of points in $[0,1]$ which remain in $[0,1]$ after one application of T_3 comprise the set
$$C_1 = \left[0, \frac{1}{3}\right] \bigcup \left[\frac{2}{3}, 1\right] = I_{11} \cup I_{12} . \tag{10}$$

2. The set of points in $[0,1]$ which remain in $[0,1]$ after two applications of T_3 comprise the set
$$C_2 = \left[0, \frac{1}{9}\right] \bigcup \left[\frac{2}{9}, \frac{1}{3}\right] \bigcup \left[\frac{2}{3}, \frac{7}{9}\right] \bigcup \left[\frac{8}{9}, 1\right] = I_{21} \cup I_{22} \cup I_{23} \cup I_{24} . \tag{11}$$

Graphically, these intervals look as follows,



These intervals look almost identical to the intervals presented earlier for the logistic map, but there is one important difference: All of the line segments that comprise a given set C_n have equal length.

The above results are easily extended to the general case: The set of points in $[0,1]$ which remain in $[0,1]$ after n applications of T_3 is the following set of 2^n closed intervals of length 3^{-n}:
$$C_n = \bigcup_{k=1}^{2^n} I_{nk} . \tag{12}$$

In other words,

$$C_n = \{x \in [0,1]\,,\ T_3^n(x) \in [0,1]\,\}\,. \tag{13}$$

From the diagram above, we see that

$$C_2 \subset C_1 \subset C_0\,. \tag{14}$$

In general,

$$C_{n+1} \subset C_n\,. \tag{15}$$

In other words, if a point x remains in [0,1] after $n+1$ iterations, which implies that $x = C_{n+1}$, then it must remain in [0,1] after only n iterations, i.e., $x =\in C_n$. But the converse does not necessarily apply: There will be points in C_n that leave after the next application of T_3.

As such, we have a nested set of **closed** intervals,

$$C_0 \supset C_1 \supset C_2 \supset \cdots\,. \tag{16}$$

We now define the set C to be the infinite intersection of this nested set, i.e.,

$$C = \bigcap_{n=0}^{\infty} C_n\,. \tag{17}$$

C is the set of points in $[0,1]$ that belong to all C_n, $n \geq 0$, i.e.,

$$C = \{x \in [0,1]\,,\ x \in C_n \text{ for all } n \geq 0\}. \tag{18}$$

Eq. (17) is stating that

$$C = \text{“}\lim_{n\to\infty} C_n\text{”}. \tag{19}$$

The set C described above is commonly known as the “ternary Cantor set” or simply “the Cantor set.”

Clearly, the set C contains the points $0, 1$, $\frac{1}{3}$ and $\frac{2}{3}$. It contains some multiples of $\frac{1}{9}$ but not all of them, i.e., it contains the points $\frac{k}{9}$ for $k \in \{0,1,2,3,6,7,8,9\}$ but not for $k\{4,5\}$. Later, we'll see that C also contains some points that you may not have guessed, e.g., $\frac{1}{4}$.

The Cantor set C, and all Cantor-like sets (we'll define this shortly), is a fascinating set. Here, we state a few basic, but very interesting properties of the set C:

1. It is bounded. This is rather obvious since $C \subset [0,1]$.

2. It is closed. This means that if $\{x_n\}_1^\infty \in C$ which is convergent, then the limit x of this sequence lies in C.

3. It is "totally disconnected." This means that if x and y are two points in C, with $x \neq y$, we can find a point z between x and y that does not lie in C.

4. It is "perfect." This means that any point $x \in C$ is the limit of a sequence of points $\{x_n\}$ which lie entirely in C.

5. It is "uncountable." A set $S \subset \mathbb{R}$ is said to be **countable** if its elements can be put into a one-to-one correspondence with the set of natural numbers, $\mathbb{N} = \{1, 2, 3, \cdots\}$. (Some texts consider $\mathbb{N}$ to be the set $\{0, 1, 2, 3, \cdots\}$. It doesn't really matter. The definition that we use is useful for the discussion below.) Essentially, a countable set S can be "counted" using the positive integers as indices. A set which is not countable (i.e., cannot be put into such a one-to-one correspondence with the set $\mathbb{N}$) is said to be **uncountable**.

Examples:

1. A finite set of elements $\{x_1, x_2, \cdots, x_N\}$ is countable. As such, the question of countability vs. uncountability is more appropriate for infinite sets.

2. The (infinite) set of real numbers $\{\frac{1}{n}\}_{n=1}^{\infty}$, i.e., $x_n = \frac{1}{n}$ for $n \geq 1$, is countable. We also can say that the set is "countably infinite.

3. The set of rational numbers in the interval $[0, 1]$ is countable. (See below.)

4. The set of real numbers in the interval $[0, 1]$ is uncountable. (See below.)

With regard to Example 3 above, here is one way to make a one-to-one correspondence between the set of rational numbers in [0,1] and the natural numbers $\mathbb{N}$. Let's call this set S, i.e., $S = \mathbb{Q} \cap [0, 1]$. First of all, $x = 0$ and $x = 1$ are rational numbers, hence members of S. We'll let these two points be the first elements of our set which will be put into one-to-one correspondence with the elements of $\mathbb{N}$.

An element $x \in S$ will have the form,

$$x = \frac{p}{q}\,, \quad \text{where} \quad p, q \in \mathbb{Z}\,, \quad 0 \leq p \leq q\,. \tag{20}$$

Let us consider $x = 0$ and $x = 1$ to be the first two elements of our set which will be put into one-to-one correspondence with the natural numbers, $\mathbb{N}$, i.e.,

$$y_1 = 0\,,\ y_2 = 1\,. \tag{21}$$

The remaining elements of S lie in the open interval $(0, 1)$. These rational numbers (all of which are positive) will have the form,

$$x = \frac{p}{q}\,, \quad \text{where} \quad p, q \in \mathbb{N}\,, \quad 1 \leq p \leq q - 1\,. \tag{22}$$

We now list all possible quotients of the above form, starting with $q = 2$ since the case $q = 1$ – for which $p = 0$ is the only possibility – has been done. For each q we consider all cases $p = 1$ to $p = q - 1$. The next few cases are:

- $q = 2$: $\frac{1}{2}$.
- $q = 3$: $\frac{1}{3}$, $\frac{2}{3}$.
- $q = 4$: $\frac{1}{4}$, $\frac{2}{4}$, $\frac{3}{4}$.

Just one more thing: You'll notice that the number $\frac{2}{4}$ in the case $q = 4$ was already generated in the case $q = 2$, so we don't need to include it in our set. As such, our algorithm will exclude any rational numbers of the form $\frac{p}{q}$ where p and q have a common factor, since they would have been generated earlier. The resulting "reordered" and "enumerated" set is then,

$$\begin{aligned} \mathbf{y} &= \{y_1, y_2, \cdots\} \\ &= \left\{0, 1, \frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \cdots\right\} . \end{aligned} \tag{23}$$

An uncountable set cannot be enumerated in this way. Such an infinite set is "larger" than a set that can be enumerated, e.g., the rational numbers. The real numbers $\mathbb{R}$ is an example of such a set. Even the set of real numbers $[0, 1]$ is uncountable. The proof of this statement is beyond the scope of this course.

What is even more fascinating is that the Cantor set C is uncountable. As we'll show later, we have essentially taken a set of length 1 out of the set $[0, 1]$ which has length 1, and the set which remains – the Cantor set C – which has length 0 (we'll explain this later) is uncountable!

Finally, the Cantor set C is an example of a **fractal** set. The term "fractal" comes from "fractional dimensional". The set C is, in some way, "thicker" than a set of points, even a countably infinite set of points, but not as "thick" as the interval of real numbers $[0, 1]$. We'll show this later.

Lecture 25

Nonlinear dynamical systems - conclusion

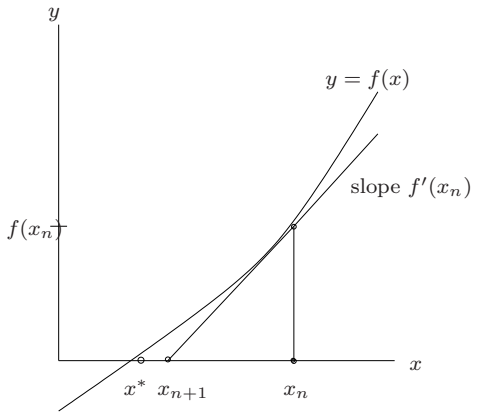
The Newton-Raphson method as a dynamical system on $\mathbb{R}$

In this section we review the well-known Newton-Raphson method, or simply "Newton's method" for finding the zeros or roots of a function $f(x)$. Our goal is to discuss Newton's method in the complex plane, in order to see some consequences of Julia-Fatou theory on the basins of attraction. Since we have encountered Newton's method on the real line earlier in this course, this first section can be skipped if desired.

The Newton iteration function associated with a function $f(x)$ is given by

$$N(x) = x - \frac{f(x)}{f'(x)}. \tag{24}$$

It is well known that iteration sequences $x_{n+1} = N(x_n)$ converge locally to a zero x^* of f, i.e. $f(x^*) = 0$. The geometric interpretation of this iteration process is sketched below. Given an iterate x_n that is "reasonably close" to the zero x^*, the iterate x_{n+1} is determined by the intersection of the tangent line to the graph of $f(x)$ at $(x_n, f(x_n))$ and the x-axis:



The points $(x_n, f(x_n))$ and $(x_{n+1}, 0)$ lie on the tangent line at $(x_n, f(x_n))$ which has slope $f'(x_n)$. Thus

$$f'(x_n) = \frac{f(x_n) - 0}{x_n - x_{n+1}}, \tag{25}$$

which is then rearranged to solve for x_{n+1}:

$$\begin{aligned} x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} && (26) \\ &= N(x_n). && (27) \end{aligned}$$

Note that zeros of f are fixed points of $N(x) : f(x^*) = 0 \Longrightarrow N(x^*) = x^*$. If fact, these are the **only** fixed points of $N(x)$, i.e.

$$\bar{x} = N(\bar{x}) \Leftrightarrow f(\bar{x}) = 0. \tag{28}$$

It is somewhat comforting to know that the Newton function has no fixed points that are not zeros of $f(x)$. This saves us from having to check such fixed points to see whether, indeed, they are zeros of $f(x)$.

However, given our study of discrete dynamical systems to date, we are naturally curious as to whether the zeros of $f(x)$ are necessarily **attractive** fixed points of $N(x)$. In other words, is it guaranteed that a fixed point $\bar{x}$ of $N(x)$ will attract nearby estimates? There would be a definite cause for concern if the answer were "No". As such, we must examine the multiplier $N'(\bar{x})$ to see whether or not $\bar{x}$ is attractive. First,

$$N'(x) = 1 - \frac{[f'(x)]^2 - f(x)f''(x)}{[f'(x)]} \tag{29}$$

$$= \frac{f(x)f''(x)}{[f'(x)]^2}. \tag{30}$$

If $\bar{x} = x^*$ is a zero of f, then it follows that

$$N'(\bar{x}) = 0. \tag{31}$$

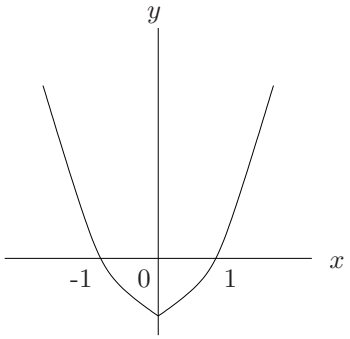
In other words, the fixed point is **attractive**.

A cautionary note should be added here: We have assumed that $f'(\bar{x}) \neq 0$. This, in fact, is equivalent to the assumption that $\bar{x} = x^*$ is a **simple zero** of $f(x)$, i.e. $f(x)$ can be factored as $f(x) = (x - x^*)g(x)$ where $g(x^*) \neq 0$. In the discussion that follows, we shall continue with this assumption.

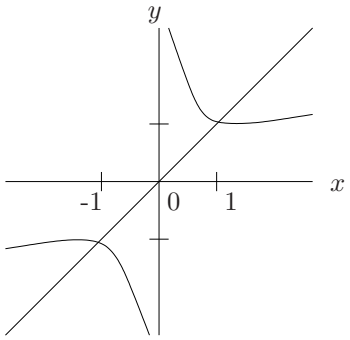
Example: $f(x) = x^2 - 1$. The zeros of f are $x_1^* = 1$ and $x_2^* = -1$. A simple calculation shows that

$$N(x) = \frac{1}{2}x + \frac{1}{2x}.$$

We sketch $f(x)$ and $N(x)$ below:



$y = x^2 - 1$



$y = \frac{1}{2}x + \frac{1}{2x}$

The reader is invited to graphically compare the Newton iteration method in both figures, using a starting point x_0 near one of the roots x_i^*.

Let us now examine the behaviour of the iteration scheme $x_{n+1} = N(x_n)$ near a fixed point x^* in a little more detail in order to get an idea of the successive errors in approximation to x^*. Since $N'(x^*) = 0$, we shall apply Taylor's theorem for $n = 1$ to $N(x)$ over an interval I containing x^*. We are also assuming that $N''(x)$ is defined for all $x \in I$. For any $x \in I$,

$$N(x) = N(x^*) + N'(x^*)(x - x^*) + \frac{1}{2!}N''(c)(x - x^*)^2, \tag{32}$$

where c lies between x and x^*. However, $N(x^*) = x^*$ and $N'(x^*) = 0$, giving

$$N(x) = x^* + \frac{1}{2}N''(c)(x - x^*)^2. \tag{33}$$

Now suppose that $|N''(x)| < M$ for all $x \in I$. Then (33) can be rearranged to give

$$|N(x) - x^*| \leq K|x - x^*|^2, \tag{34}$$

where $K = \frac{1}{2}M$. If x is to be considered an approximation to x^*, then the absolute error of this approximation is given by $\epsilon = |x - x^*|$. Eq. (34) tells us that the error in approximating x^* by $N(x)$ is

$$|N(x) - x^*| \leq K\epsilon^2. \tag{35}$$

This is known as the **quadratic convergence** of Newton's method. Quadratic convergence is due to the fact that $N'(x^*) = 0$. From (35), the convergence of the iteration sequence $x_{n+1} = N(x_n)$ to a fixed point $\bar{x} = x^*$ of $N(x)$ is given by

$$|x_{n+1} - x^*| \leq K\,|x_n - x^*|^2. \tag{36}$$

For example, if $\epsilon_n = |x_n - x^*| \cong 10^{-5}$, then we expect $\epsilon_{n+1} = |x_{n+1} - x^*|$ to be on the order of $10^{-9} - 10^{-10}$, roughly a doubling of the number of digits of accuracy in approximating x^*.

Example: We return to $f(x) = x^2 - 1$ with Newton function

$$N(x) = \frac{1}{2}x + \frac{1}{2x}.$$

If we use the starting value $x_0 = 10$, then the following iterates are produced, to ten digits of accuracy:

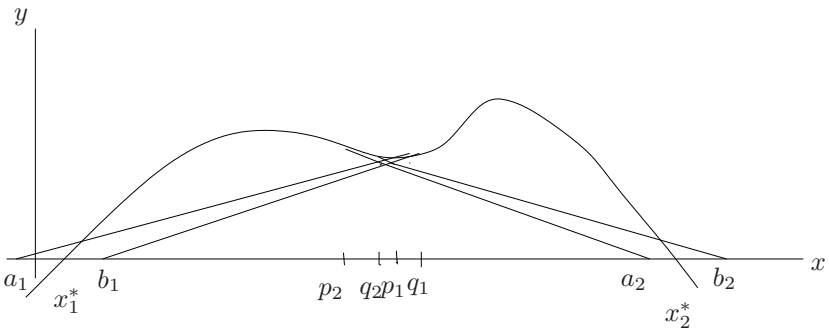
n	x_n
1	5.050
2	2.624009901
3	1.502553013
4	1.084043468
5	1.003257851
6	1.000005290
7	1.000000000

The rapid convergence of the iterates is clearly demonstrated.

Exercise: Construct the Newton function $N(x)$ associated with the quadratic $f(x) = x^2 - a^2$, $a > 0$. sketch $f(x)$, $N(x)$ and determine the basins of attraction of the roots $x_1^* = a$, $x_2^* = -a$.

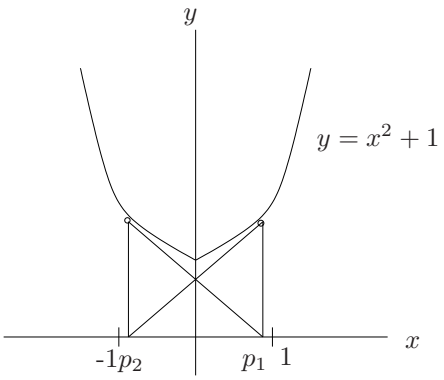
In general, the basins of attraction for roots of a function $f(x)$ will not be so "nice". Indeed, Newton's method was constructed with the idea of *local convergence*, i.e. starting with a reasonable approximation x_0 of a root x^* and then constructing a sequence $x_{n+1} = N(x_n)$ of improved estimates to x^*. This must always be kept in mind in applications.

The geometric picture of Newton's method can give us an idea of how the basins of attraction for a root can be more complicated in structure. For example, consider the situation sketched below.



From the graph, it seems reasonable that points $x \in [a_1, b_1]$ belong to the basin of attraction $W(x_1^*)$ and points $x \in [a_2, b_2]$ belong to $W(x_2^*)$. However, it also appears that points $x \in [p_1, q_1]$ are mapped by $N(x)$ into the interval $[a_1, b_1]$, so that $[p_1, q_1] \subset W(x_1^*)$. Likewise $[p_2, q_2] \subset W(x_2^*)$. Indeed, depending upon the behaviour of the graph of $f(x)$ for $x < a_2$ and $x > b_1$, it is possible that points lying in either of the intervals (b_1, p_2) or (q_1, a_2) are eventually mapped into $[a_1, b_1]$ or $[a_2, b_2]$.

As well, it is also possible – in fact necessary – that there exist points $x \in \mathbb{R}$ that do not belong to either of the basins of attraction $W(x_1^*)$ or $W(x_2^*)$. These points would include "boundary points", i.e. those points that separate the sets $W(x_1^*)$ and $W_2(x_2^*)$. In order to obtain an idea of the complicated behaviour that may exist in a region with a local minimum such as the region containing $[p_1, q_1]$ and $[p_2, q_2]$ above, let us examine the behaviour of Newton's method applied to the function $f(x) = x^2 + 1$. Of course, $f(x)$ has no real roots, so $N(x)$ has no real fixed points. Nevertheless, there is some interesting dynamical behaviour here.



The Newton's function for $f(x) = x^2 + 1$ is

$$N(x) = x - \frac{x^2+1}{2x} = \frac{1}{2}x - \frac{1}{2x}\,. \tag{37}$$

It turns out that $N(x)$ has a two-cycle (p_1, p_2), i.e. $N(p_1) = p_2$ and $N(p_2) = p_1$, as shown above.

Exercise: Show that $(p_1, p_2) = \left(\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}\right)$.

The next natural question is, "Is this two-cycle attractive or repulsive?"

Exercise: Show that $|N'(p_1)N'(p_2)| = 4$, implying that the two-cycle is repulsive.

It turns out that $N(x)$ possesses a great number of cycles all of which are repulsive. In fact, $N(x)$ defines a chaotic dynamical system on $\mathbb{R}$. However, this is beyond the scope of this discussion.

Newton's Method in the complex plane $\mathbb{C}$

In this section we examine Newton's method for finding roots of functions in the complex plane, i.e. given a complex-valued function f of a complex variable $z \in \mathbb{C}$, we define $N : \mathbb{C} \to \mathbb{C}$ as

$$N(z) = z - \frac{f(z)}{f'(z)}. \tag{38}$$

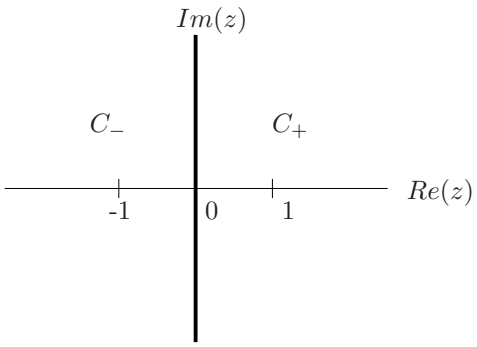
As we show only briefly below, Newton's method in the complex plane demonstrates some fascinating dynamics. If anything, the few results shown below should serve as a motivation to explore the subject of iterated complex mappings even further.

Square roots of unity: $f(z) = z^2 - 1$

We return to our original motivating example, $f(z) = z^2 - 1$, with roots $z_1^* = 1$ and $z_2^* = -1$, where $z \in \mathbb{C}$. We are concerned with the question: "What are the basins of attraction of these two roots, i.e. $W(1)$ and $W(-1)$?" Recall that in the real-variable case, $W(1) = \{x \in \mathbb{R} \,|\, x > 0\}$ and $W(-1) = \{x \in \mathbb{R} \mid x < 0\}$. One might expect, on the basis of symmetry, that these basins of attraction are "continued" into the complex plane to produce the complex basins of attraction,

$$W(1) = \{z \in \mathbb{C} \,|\, Re(z) > 0\} \quad \text{(right half-complex plane} \quad \mathbb{C}_+) \tag{39}$$

$$W(-1) = \{z \in \mathbb{C} \,|\, Re(z) < 0\} \quad \text{(left half-complex plane} \quad \mathbb{C}_-). \tag{40}$$



We shall show that, indeed, this is the case. This result was established by the mathematician A. Cayley in 1879.

The imaginary axis $\text{Im}(z) = \{z \in \mathbb{C} \,|\, z = ti,\, t \in \mathbb{R}\}$ serves as a boundary between the right and left half-planes, $\mathbb{C}_+$ and $\mathbb{C}_-$, respectively. We now examine the action of the Newton function associated with $f(z) = z^2 - 1$, namely,

$$N(z) = \frac{1}{2}z + \frac{1}{2z}, \tag{41}$$

on the imaginary axis $\mathrm{Im}(z)$. If we set $z = ti$, then, for $t \neq 0$,

$$\begin{aligned} N(z) = N(ti) &= \frac{t}{2}i + \frac{1}{2ti} \\ &= \frac{1}{2}\left[t - \frac{1}{t}\right] i. \end{aligned} \tag{42}$$

Since $t \in \mathbb{R}$, it follows that $N(ti) \in \mathrm{Im}(z)$. This implies that the imaginary axis $\mathrm{Im}(z)$ is invariant under the action of N, i.e. $N : \mathrm{Im}(z) \to \mathrm{Im}(z)$. This is somewhat encouraging if we want $\mathrm{Im}(z)$ to be a boundary curve. (Note, however, that the real axis, $\mathrm{Re}(z)$, is also necessarily invariant under the action of $N(z)$, since the coefficients of $N(z)$ in (41) are real.) However, we must rule out the possibility that points may jump from $\mathbb{C}_+$ to $\mathbb{C}_-$, or vice versa, under the action of N.

Let $z = a + bi$, $a, b \in \mathbb{R}$. Then, from (41),

$$N(z) = N(a+bi) = \frac{1}{2}(a+bi) + \frac{1}{2}\frac{1}{a+bi} \tag{43}$$

$$= \frac{1}{2}(a+bi) + \frac{1}{2}\frac{a-bi}{a^2+b^2} \tag{44}$$

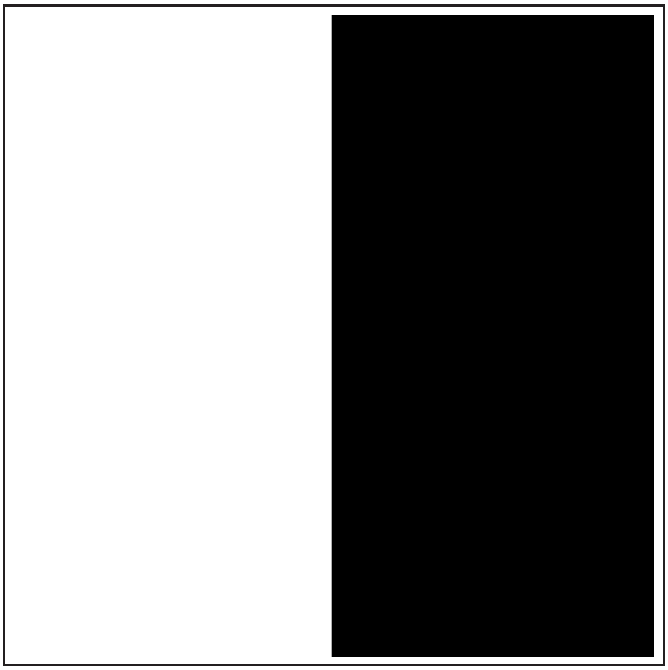
$$= \frac{1}{2}a\left[1 + \frac{1}{a^2+b^2}\right] + \frac{1}{2}b\left[1 - \frac{1}{a^2+b^2}\right] i. \tag{45}$$

It follows that if $a = \mathrm{Re}(z) > 0$, then $\mathrm{Re}(N(z)) > 0$. As well, if $a = \mathrm{Re}(z) < 0$, then $\mathrm{Re}(N(z)) < 0$. Therefore, we may conclude that $N : \mathbb{C}_+ \to \mathbb{C}_+$ and $N : \mathbb{C}_- \to \mathbb{C}_-$, so that points cannot hop from $\mathbb{C}_+$ to $\mathbb{C}_-$ or vice versa.

So far, we have established that $\mathbb{C}_+$ (right half-plane), $\mathbb{C}_-$ (left half-plane) and $\mathrm{Im}(z)$ (imaginary axis) are invariant under the action of $N(z)$. There is still some work to do to show that, in fact, $W(1) = \mathbb{C}_+$ and $W(-1) = \mathbb{C}_-$. However, since such work is beyond the scope of this course, we merely accept the result and refer the reader to the literature for further details.

A look at Eq. (42) will show that the action of N on $\mathrm{Im}(z)$ is equivalent to the action of Newton's map for $f(x) = x^2 + 1$ on the real line as we saw back in Eq. (37). This should not be too surprising when one realizes that the complex function $f(z) = z^2 + 1$ has purely imaginary roots $z_1^* = i$ and $z_2^* = -i$. Here, the boundary between basins of attraction $W(i)$ and $W(-i)$ is the real line $\mathbb{R}$.

In the figure below is shown a computer approximation of the basins of attraction of Newton's method applied to $f(z) = z^2 - 1$ over the square region $-1 \leq \mathrm{Re}(z) \leq 1$,$-1 \leq \mathrm{Im}(z) \leq 1$. The region in black depicts the basin of attraction $W(1)$ of the root $z_1^* = 1$. The region in white depicts the basin of attraction $W(2)$ of the root $z_2^* = -1$. (Details on how the plot was constructed will be provided later.) The plot is in accordance with our conjecture.



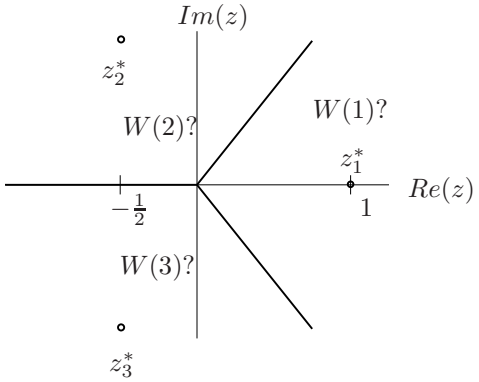
Basins of attraction of the two square roots of unity for Newton's method applied to the function $f(z) = z^2 - 1$ in the region $-1 \leq \mathrm{Re}(z) \leq 1$, $-1 \leq \mathrm{Im}(z) \leq 1$ of the complex plane $\mathbb{C}$.

Cube roots of unity: $f(z) = z^3 - 1$

Let us now consider Newton's method applied to the function $f(z) = z^3 - 1$, with roots, $z_1^* = 1$, $z_2^* = -\frac{1}{2} + \frac{1}{2}\sqrt{3}i$, $z_3^* = -\frac{1}{2} - \frac{1}{2}\sqrt{3}i$, i.e. the cube roots of unity. The Newton function associated with $f(z)$ is

$$N(z) = z - \frac{z^3 - 1}{3z^2} = \frac{2z}{3} + \frac{1}{3z^2}\,. \tag{46}$$

On the basis of the previous discussion, one might expect that the basins of attraction for these roots would be the appropriate one-third sectors of the complex plane $\mathbb{C}$ whose boundaries are rays emanating from the origin and perpendicularly bisecting the line segments joining the roots z_i^*, as shown in the diagram below.



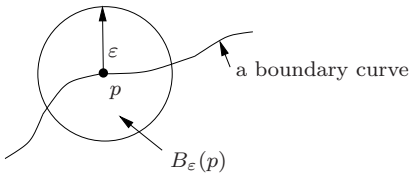
In fact, Cayley posed this question in his 1859 paper, which remained unanswered until the works of

the French mathematicians G. Julia and P. Fatou in the early 1900's. Julia and Fatou were responsible for the development of the theory of iteration of rational functions in the complex plane, a subject which has received a great deal of renewed interest over the past 15-20 years, both with regard to theory as well as applications.

In fact, the Julia-Fatou theory revealed that the picture of basin boundaries drawn on the previous page could **not** be correct, since the basin boundaries would have to exhibit a most remarkable property that we now briefly describe. Let $p \in \mathbb{C}$ be a point on a basin boundary curve associated with Newton's method for $f(z) = z^3 - 1$, with the three attractive fixed points z_i^*. Now let $B_\epsilon(p)$ denote a ball of radius $\epsilon > 0$ centred at p, i.e.

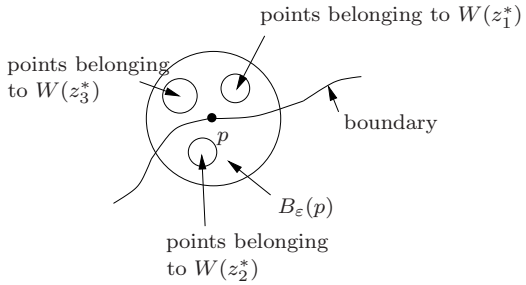
$$B_\epsilon(p) = \left\{ z \in \mathbb{C} \,\middle|\, |z - p| < \epsilon \right\}.$$

This is sketched below:

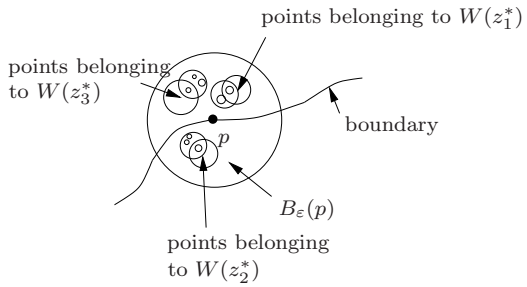


We emphasize that ϵ can be **any** positive number, however small in magnitude.

The remarkable result of Julia-Fatou theory is that the ball $B_\epsilon(p)$ **must** contain points that belong to **each** of the basins attraction $W(z_1^*)$, $W(z_2^*)$ and $W(z_3^*)$! In other words, we must be able to find points in $B_\epsilon(p)$ that belong to $W(z_1^*)$, i.e. points that, under the action of $N(z)$, eventually get closer and closer to the root z_1^*, as well as points that belong to $W(z_2^*)$, and points that belong to $W(z_3^*)$. Let us try to sketch this situation below:



But wait a minute! The set of points belonging to $W(z_1^*)$ must have a boundary curve that separates this set from the set of points belonging to $W(z_2^*)$ which, in turn, must have a boundary curve that separates it from points in $W(z_3^*)$. Therefore, we must be able to pick **any** points from these boundary curves and draw little "epsilon-balls" around them that contain points from each of the three basins of attraction:



The reader may begin to appreciate the immense complication of such a boundary curve. Essentially, the three basins of attraction $W(z_1^*)$, $W(z_2^*)$ and $W(z_3^*)$ **must share a common boundary curve**. The implications in the case of three basins of attraction are enormous – the boundary curve must have a complicated, self-similar, "fractal" structure.

In the early 1900's and the absence of computers, it was difficult for mathematicians to envisage the structure of these basins of attraction and their complicated boundary curves. Today, with computers, we can at least obtain an idea of the complexity of these sets and curves. A computer approximation of the basins of attraction for Newton's method applied to $f(z) = z^3 - 1$ is presented in the figure on the next page.

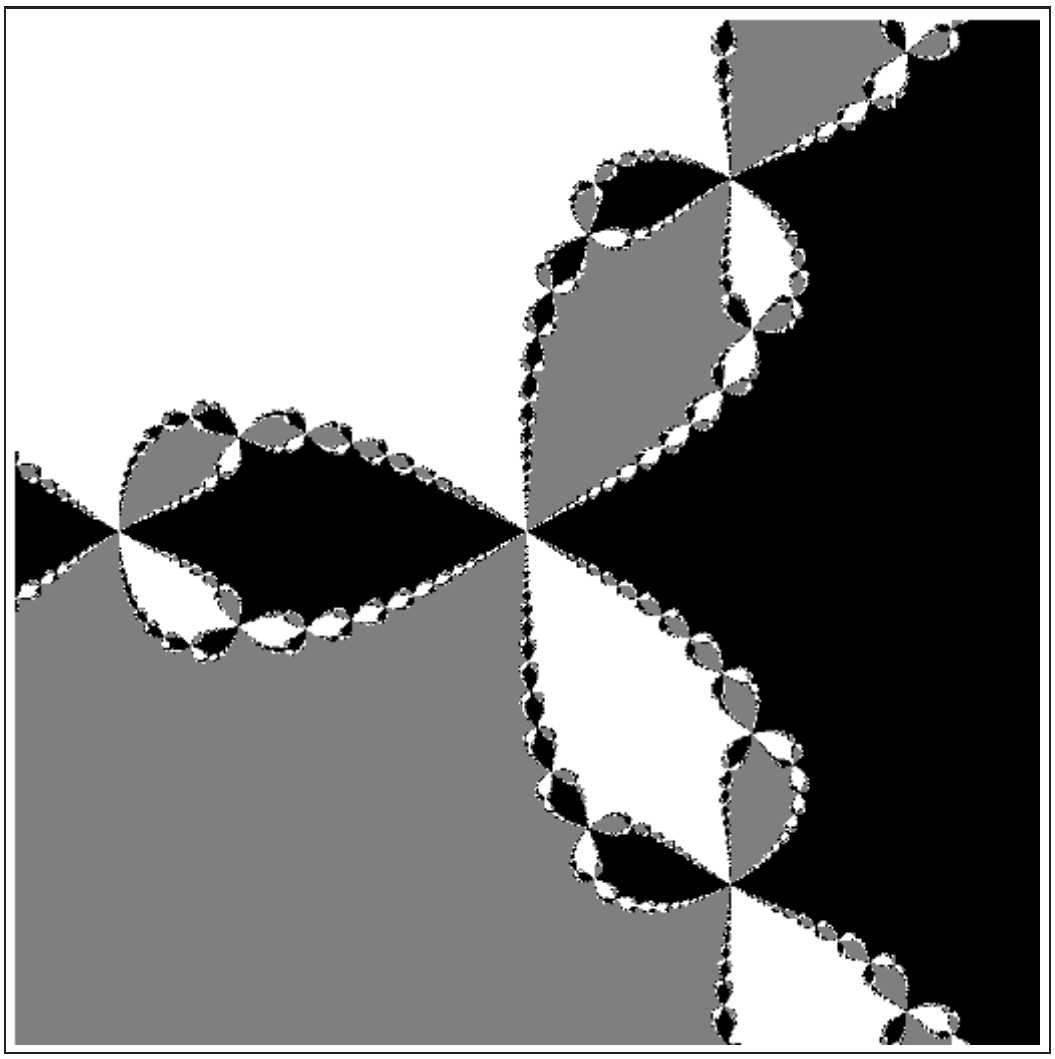
We outline briefly how this picture was obtained. The plot represents the square region $-1 \leq \mathrm{Re}(z) \leq 1$, $-1 \leq \mathrm{Im}(z) \leq 1$, divided up into a 512×512 pixel grid. The geometric centre of each pixel p_{ij}, $1 \leq i, j \leq 512$, determined and then used as a starting point z_0 for the Newton iteration method (cf. Eq. (46)),

$$z_{n+1} = \frac{2}{3} z_n + \frac{1}{3z_n^2}, \qquad n \geq 0.$$

For each n, the term z_n is checked for proximity to each of the roots z_1^*, z_2^* and z_3^*. If $|z_n - z_i^*| < \epsilon$, where $\epsilon > 0$ is a preassigned value (say, 0.5), then the point z_0 is classified as belonging to the basin of attraction $W(z_i^*)$ and the pixel p_{ij} is shaded accordingly (white if $i = 1$, gray if $i = 2$, black if $i = 3$).

Note that it is virtually impossible for a point z_0/pixel p_{ij} to be an element of the boundary curve of the basins of attraction. As in the case of the Cantor sets associated with the logistic maps $f_a(X) = ax(1 - x)$, $a > 4$, the boundary curves are **repeller** sets, composed of repulsive n-cycle points. As such, in order to remain on such curves, we would have to know the points on these curves to infinite numerical precision, which is impossible. As a result, our pixel points will almost certainly belong to one of the three basins of attraction. The boundary curve is then implicitly defined by any change in shading between a pixel and its neighbouring pixel.

In the figure below is shown a magnification of the above figure near the origin – the region $-0.25 \leq \mathrm{Re}(z) \leq 0.25$, $-0.25 \leq \mathrm{Im}(z) \leq 0.25$. Similar structures as in the previous figure are observed. Here we simply state that if one magnifies, to any degree, any region that contains boundary points, i.e., points from the Julia set J of the Newton function, similar structures will be observed. This attests to the infinite "self-similar" nature of the Julia set boundary.

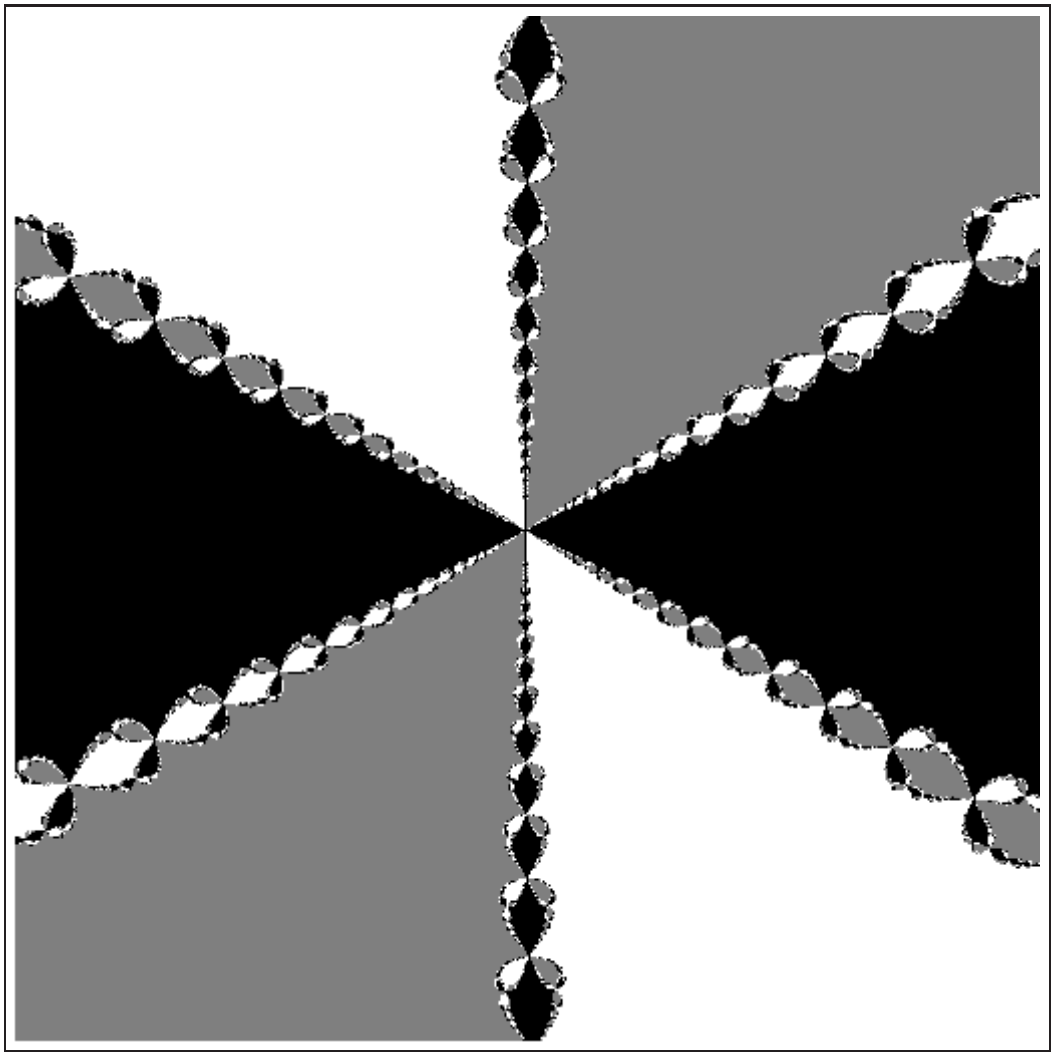


Basins of attraction of the three cube roots of unity for Newton's method applied to the function $f(z) = z^3 - 1$ in the region $-1 \leq \mathrm{Re}(z) \leq 1$, $-1 \leq \mathrm{Im}(z) \leq 1$ of the complex plane $\mathbb{C}$.

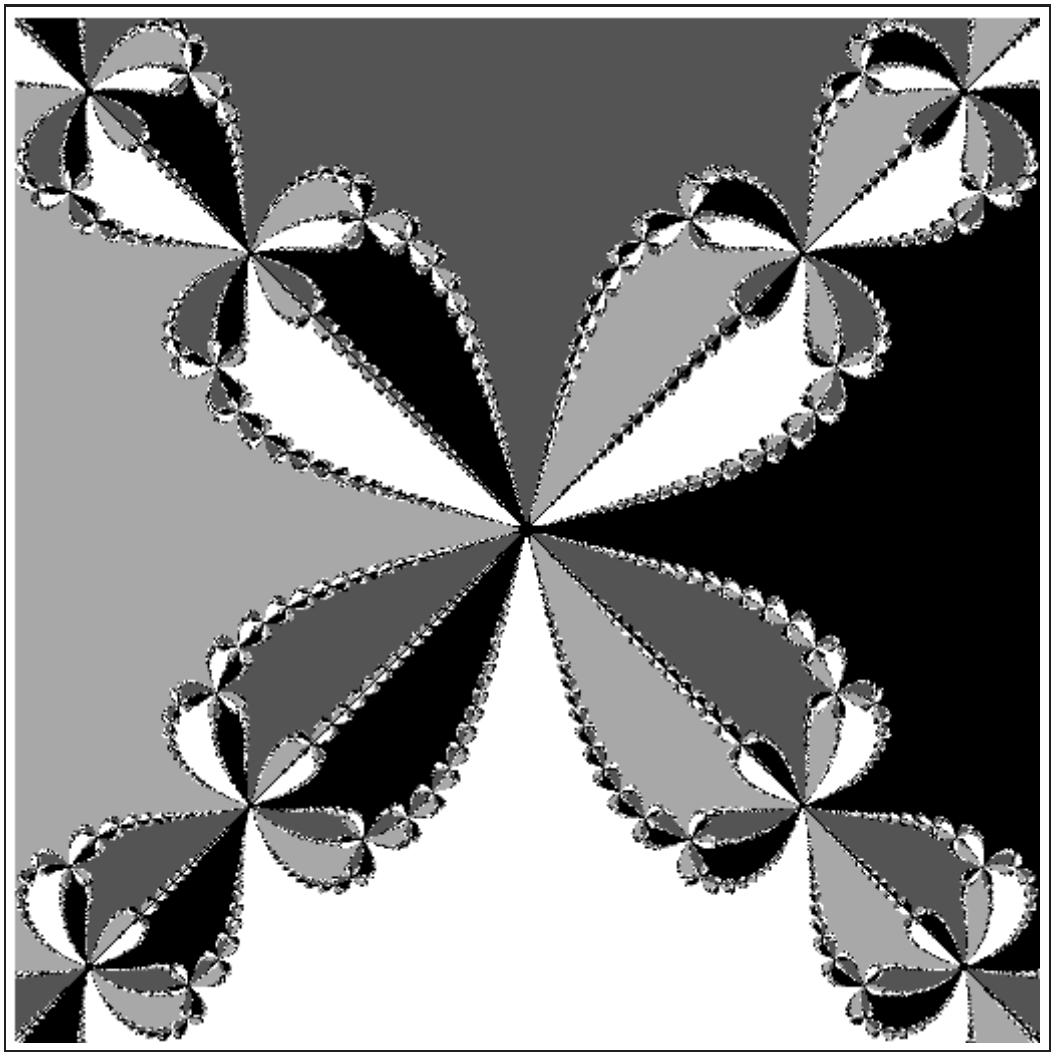
Fourth roots of unity: $f(z) = z^4 - 1$

The figure below shows the basin boundaries for the fourth roots of unity, $1, -i, -1$ and i, when Newton's method is applied to $f(z) = z^4 - 1$. Once again, the plot represents the square region $-1 \leq \mathrm{Re}(z) \leq 1$, $-1 \leq \mathrm{Im}(z) \leq 1$.

Finally, note that Julia-Fatou theory allows the existence of a smooth basin boundary for the case of two roots, i.e. $f(z) = z^2 - 1$, namely, the imaginary axis $\mathrm{Im}(z)$. Here we state, without proof, that the imaginary axis $\mathrm{Im}(z)$ is the Julia set of the Newton function $N(x)$ in Eq. (41) associated with $f(z) = z^2 - 1$. (As such, it contains all repulsive periodic points of $N(x)$.) For any point $p \in \mathrm{Im}(z)$, any ball $B_\epsilon(p)$ around p trivially contains points from $\mathbb{C}_+$ and $\mathbb{C}_-$, the basins of attraction of $z_1^* = 1$ and $z_2^* = -1$.



Basins of attraction of the three cube roots of unity for Newton's method applied to the function $f(z) = z^3 - 1$ in the region $-0.25 \leq \mathrm{Re}(z) \leq 0.25$, $-0.25 \leq \mathrm{Im}(z) \leq 0.25$ of the complex plane $\mathbb{C}$.



Basins of attraction of the four fourth roots of unity for Newton's method applied to the function $f(z) = z^4 - 1$ in the region $-1 \leq \mathrm{Re}(z) \leq 1$, $-1 \leq \mathrm{Im}(z) \leq 1$ of the complex plane $\mathbb{C}$.

Lecture 26

An introduction to fractals and fractal geometry

A motivating example – the length of boundaries and coastlines

Recall the following problem from Calculus: Given a bounded, continuous curve $C \in \mathbb{R}^2$, find, if possible, its length.

If the curve is **rectifiable**, i.e., it is piecewise "smooth" (piecewise C^1 or continuously differentiable), then its length is finite. If you have formulas for C, e.g., parametrization of C, you may be able to compute its length L.

A more practical problem: Suppose that you have a curve C – perhaps a boundary curve of part of a region D in the plane, for example, the boundary that you share with a neighbour. If the boundary is straight, or at least in principle, straight, then there should be no problem. Just hammer a stake into the ground at each end of the curve, tie the end of a string/rope around one stake, go to the other stake, pull the string/rope until taut, tie the string/rope to the stake. Then measure the length of the (hopefully) straight rope.

But what if the boundary curve is not straight, for whatever reason(s)?

Practical solution: Use a measuring apparatus (e.g., a string) of a sufficiently small* length ℓ_1, yet not too small so that it is still practical, e.g., $\ell_1 = 10$ m. Then start at one end of C and connect points on C with the taut string. If you need to do this N_1 times before you reach the other end, your approximation is

$$L \approx L_1 = N_1 \ell_1 \,. \tag{47}$$

***A note on "sufficiently small":** You don't want ℓ_1 to be too long so that important details are not detected, e.g., small "nooks" or "crannies."

But suppose that you desire a better approximation to L. You would most probably consider using a measuring apparatus of smaller length $\ell_2 < \ell_1$, say 1 m. Of course, this is less convenient, but you may be happier with the result – a better approximation,

$$L \approx L_2 = N_2 \ell_2 \,. \tag{48}$$

And if you're not satisfied with that result, you can try an apparatus of even smaller length, i.e., $\ell_3 < \ell_2$, to arrive at a hopefully better approximation,

$$L \approx L_3 = N_3 \ell_3 \,. \tag{49}$$

If you had the time and energy to use measuring devices of even smaller length, say,

$$\ell_1 > \ell_2 > \ell_3 > \cdots , \tag{50}$$

you might find that for some sufficiently large n, your estimates L_n and L_{n+1} may be within some prescribed tolerance, say $\delta > 0$, i.e.,

$$|L_n - L_{n+1}| < \delta , \tag{51}$$

and you would be comfortable in concluding that the length of your boundary curve C is

$$L \approx L_n \quad \text{or perhaps} \quad L_n \pm \delta . \tag{52}$$

Let's just move this practical example back into the more theoretical mathematical world, and assume that your boundary curve $C \in \mathbb{R}^2$ is rectifiable. In this case, if you used a set of measuring devices with monotonically decreasing lengths, i.e.,

$$\ell_{n+1} < \ell_n \tag{53}$$

along with the condition that

$$\lim_{n\to\infty} \ell_n = 0 . \tag{54}$$

In this case, then we expect the sequence of estimated lengths,

$$L_n = N_n \ell_n , \tag{55}$$

to converge to L as $n \to \infty$, i.e.,

$$\lim_{n\to\infty} L_n = L . \tag{56}$$

At this point, let's use the generic "ϵ" to denote the length of a very small measuring stick and consider the case that $\epsilon \to 0^+$. We'll let $N(\epsilon)$ denote the number of measuring sticks of length ϵ to "cover" curve C, i.e., to get from one end to the other, so that

$$L(\epsilon) = N(\epsilon)\epsilon \cong L \tag{57}$$

is the approximation to the true length L of the curve C.

We also note that the number of measuring sticks of length ϵ needed to "cover" C is roughly

$$N(\epsilon) \cong L\epsilon^{-1} \quad \text{as} \quad \epsilon \to 0 . \tag{58}$$

This seemingly trivial result will be important in future discussions. In fact, the most important part of (58) is the exponent associated with the ϵ. We shall be content with the statement,

$$N(\epsilon) = A\epsilon^{-1} \quad \text{as} \quad \epsilon \to 0 . \tag{59}$$

or simply,

$$N(\epsilon) = O(\epsilon^{-1}) \quad \text{as} \quad \epsilon \to 0\,. \tag{60}$$

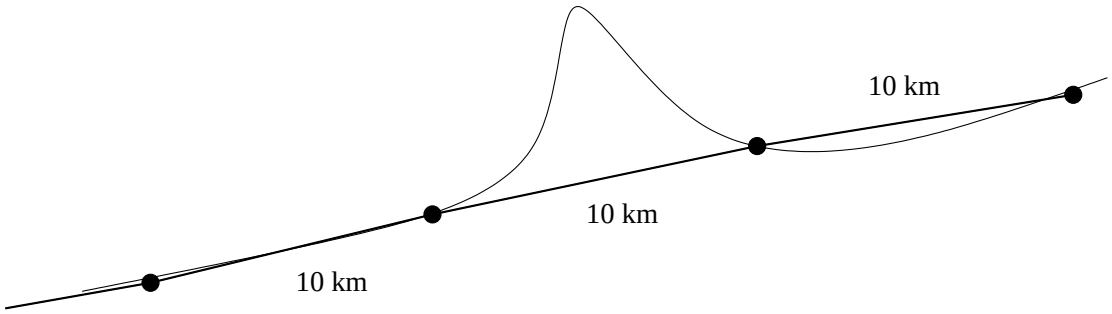
But what if our boundary curve C were quite "irregular"? And what about "irregular curves" that encountered in nature? For example, let us consider the classic **Coastline Problem** discussed by Benoit Mandelbrot in his book, *Fractal Geometry of Nature*. Suppose you wanted to measure, or at least assign a length to, the coastline of an island, say Great Britain. We'll define the coastline as the hypothetical curve of intersection of the surface of the perfectly calm sea with the landform.

Once again, we can imagine using "measuring sticks" of decreasing length ℓ_n, $n = 1, 2, 3, \cdots$, i.e.,

$$\ell_1 > \ell_2 > \cdots > \ell_n > \ell_{n+1} \cdots\,. \tag{61}$$

Perhaps we start with $\ell_1 = 100$ km, and simply use a map to estimate the length of the coastline. But if we're serious, and willing to do the work, perhaps we could consider a measuring "stick" of length $\ell_1 = 10$ km and arrive at a length that would seem to be satisfactory.

But a 10 km measuring stick could completely ignore a bay or inlet, and almost certainly the entrance of a river into the sea, as sketched below.



And by the way, what do we do with regard to rivers? Do we proceed to measure the length of the landform/water surface interface all the way up the river? Probably not – that could easily increase our length estimate many times. We may have to make some kind of artificial decision. For ease of argument, let's assume that we can make some kind of decision here:

- A 1 km measuring stick may be able to take large bays or inlets into some consideration.
- But a 100 m measuring stick would do a much better job here.
- A 1 m stick would do even better, but there would still be crevices or large rocks that could be entirely ignored.
- A 10 cm stick would do better yet, and a 1 cm stick even better.

- A 1 mm stick would pick up fluctuations due to gravel.

At this point, we can well imagine that our estimates,

$$L_n = N(\ell_n)\ell_n \,, \tag{62}$$

of the length L of the coastline will not only be increasing as we pick up more details of the coastline curve, but will most probably not be appearing to converge to a limiting value L.

And what if we continue this procedure into to the microscopic domain? And finally into the atomic domain? Of course, this sounds rather silly in a supposedly real-world application. But we shall be letting $\ell_n = \epsilon \to 0$ in our mathematical discussions.

But back to the "real world," Mandelbrot cites the work of Louis Fry Richardson (1881-1953), a British scientist who was interested in a variety of natural studies, including meteorology and fluid dynamics as well as the more general problem of measurement. (He was also interested in the problem of analyzing war mathematically.) With regard to the latter, he found, for example, significant differences in reported measured values of boundaries, e.g.,

- Spain-Portugal border: 987 vs. 1214 km
- Netherlands-Belgium border: 380 vs. 449 km

Plate 33 of Mandelbrot's book, presented on the next page of these notes, reproduces, according to Mandelbrot, "Richardson's experimental measurement of length performed on various curves using equal-sized polygons of increasingly short side ϵ." It is not clear from Mandelbrot's discussion whether or not Richardson actually estimated the lengths of the coastlines and boundaries in the real-world case – the use of polygons, however, suggests that he performed the analysis on maps. The figure below shows such an estimation of the length of the coast of Great Britain using measuring sticks of 200, 100 and 50 km.

As Mandelbrot writes below Plate 33, the remarkable discovery is that in the case of coastlines, the approximate lengths $L(\epsilon)$ do not stabilize. As the length ϵ of the measuring stick tends to zero, the approximate lengths $L(\epsilon)$, as plotted on doubly logarithmic paper, "fall on a straight line of negative slope. This is true of boundaries between countries." This is a consequence of the experimental fact that the number, $N(\epsilon)$, of measuring sticks of length ϵ needed to cover the curve, grows as

$$N(\epsilon) \cong A\epsilon^{-D} \quad \text{as} \quad \epsilon \to 0 \,, \tag{63}$$

where D is a real number, and not an integer, greater than 1. As a result, the estimated length $L(\epsilon)$ is given by

$$L(\epsilon) \cong N(\epsilon)\,\epsilon \cong A\epsilon^{1-D} \,. \tag{64}$$

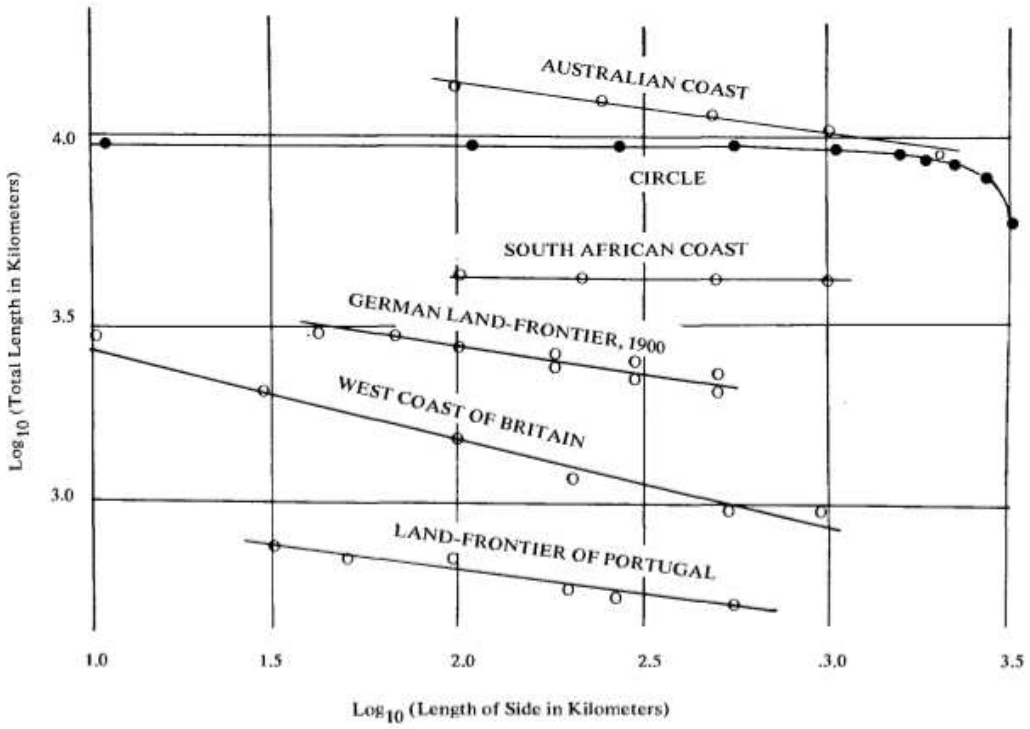


Plate 33 ¤ RICHARDSON'S EMPIRICAL DATA CONCERNING THE RATE OF INCREASE OF COASTLINES' LENGTHS

This Figure reproduces Richardson's experimental measurements of length performed on various curves using equal-sided polygons of increasingly short side ϵ. As expected, increasingly precise measurements made on a circle stabilize very rapidly near a well-determined value.

In the case of coastlines, on the contrary, the approximate lengths do not stabilize at all. As the yardstick length ϵ tends to zero, the approximate lengths, as plotted on doubly logarithmic paper, fall on a straight line of negative slope. The same is true of boundaries between countries. Richardson's search in encyclopedias reveals notable differences in the lengths of the common land frontiers claimed by Spain and Portugal (987 versus 1214 km), and by the Netherlands and Belgium (380 versus 449 km). With a slope of –0.25, the 20% differences between these claims can be accounted for by assuming that the ϵ's differ by a factor of 2, which is not unlikely.

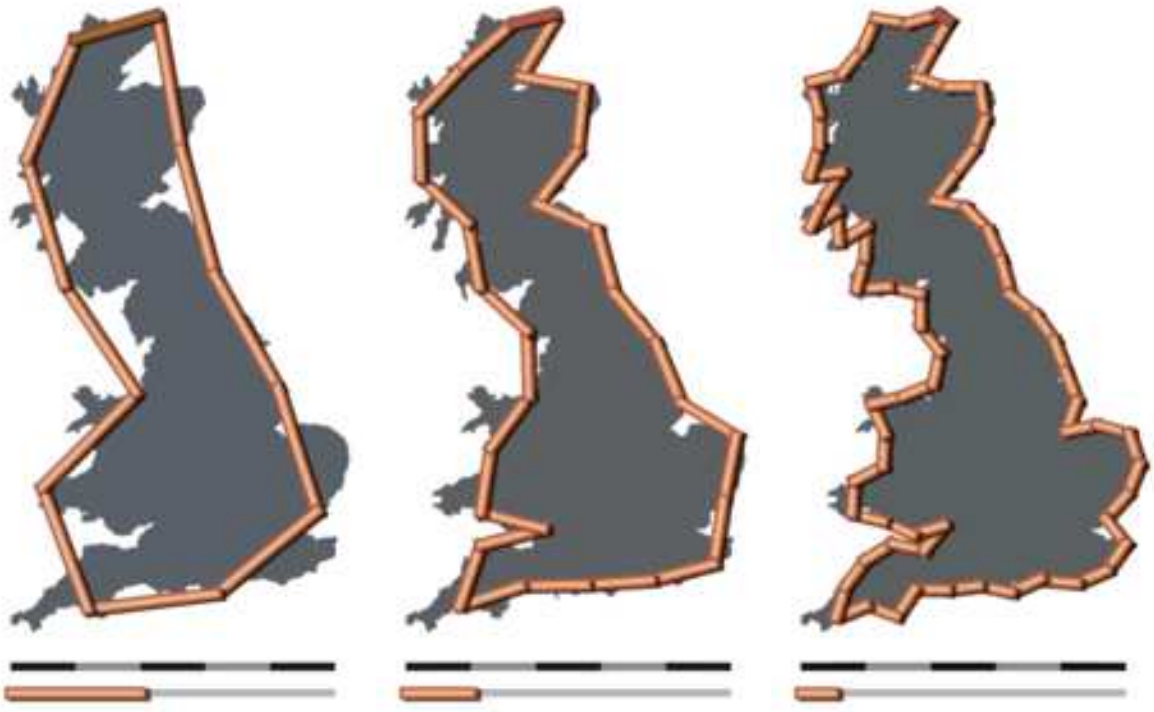
To Richardson, his lines' slopes had no theoretical interpretation. The present Essay, on the other hand, interprets coastlines as approximate fractal curves, and uses the slope of each line as an estimate of 1–D, where D is the fractal dimension. ■

33

Taking logarithms of both sides (Richardson used base-10 logarithms),

$$\log L \cong (1 - D) \log \epsilon + \log A \,. \tag{65}$$

As such, a log L vs. log ϵ plot should yield straight lines with slopes $1 - D \leq 0$. The different negative slopes observed in Plate 33 result from different D values for each boundary. In his 1967 paper, "How



Estimating the length of the coastline of Great Britain using 200 km (left), 100 km (center) and 50 km (right) measuring sticks.

Long Is the Coast of Britain?" Mandelbrot proposed that Richardson's exponent, D, be interpreted as a "fractal dimension," a term that we shall explain in the next lecture. Greater D values imply greater irregularity, i.e., a greater departure from smoothness. We shall discuss the idea of fractal dimension in the next lecture.

Note that Plate 33 of Mandelbrot's book shows that for rectifiable curves, such as a circle, the length estimates $L(\epsilon)$ do stabilize as $\epsilon \to 0$: The limiting zero-slope of the curve indicates that $1 - D = 0$, implying that $D = 1$, the so-called **topological dimension** of a line or curve. The more negative the slope $1 - D$ is, the higher the value of $D > 1$.

Richardson's analysis suggests that the discrepancies in the lengths of borders reported above is due to the use of measuring sticks of different sizes in each problem.

Moral of the story: For "irregular" curves, the higher the resolution that you use to measure their lengths, i.e., the smaller the measuring sticks, the more detail you measure, which leads to longer and longer length estimates. And the appearance of new detail at smaller scales does not seem to stop – it keeps appearing at smaller and smaller scales, leading to an apparent "blowup" of the length estimate.

Postscript: In reality, you can't go to measuring sticks of zero length. Practically, we can characterize the behaviour of length/area/volume estimates over a finite range of small scales. Over these scales, there can be a non-integral or fractal scaling exponent D and it does have repercussions on the physical properties of materials being examined. More on this later.

Recall that we have already encountered some examples of "fractal sets" in earlier sections: (1) the Cantor and Cantor-like sets of Section 3 that were produced by the repeated removal of open sets from closed intervals on the line and (2) the complicated boundaries of basins of attraction for Newton's method in the complex plane. Fractal sets are generally complicated sets that exhibit some kind of "self-similarity:" The magnification of any portion of a fractal set S looks quite similar to the original set S. Repeated magnifications of portions of portions of these sets will look similar as well. The reader may already have developed an appreciation of this property from our earlier look at the Cantor set.

There are a number of definitions of "fractal sets," all of which involve concepts from real analysis that extend beyond the scope of this course. Nevertheless, it is possible to formulate a reasonable working definition of fractal sets that does not differ too much from the technical definitions. As you will see below, the idea of "sets of non-integer dimension" will naturally follow from the problem of trying to find the "sizes" of complicated sets such as those mentioned above.

The treatment of the subject in this section is quite limited, both in terms of the mathematics as well as examples and applications. The reader interested in seeing more of these details is advised to look at the book, *Fractals Everywhere*, by Michael Barnsley. (This book is directed at senior undergraduates.) For a more pictorial treatment, along with many historical perspectives, the celebrated work, *The Fractal Geometry of Nature*, by Benoit Mandelbrot, is highly recommended. This is the book that essentially brought fractal geometry and the results of earlier mathematicians to the attention of the "outside world" of scientists.

Measuring the sizes of "Euclidean" Sets in $\mathbb{R}^k$

In what follows, we shall be discussing the idea of measuring the sizes of sets S of points in $\mathbb{R}^k$, with $k = 1, 2, 3$, in accordance with our usual experience of dealing with (1) subsets of the real line $\mathbb{R}$, (2) sets in the plane $\mathbb{R}^2$ or (3) sets in three-dimensional space $\mathbb{R}^3$.

Since childhood, we have become accustomed to the idea of measuring sizes of "Euclidean" objects: lines or curves are measured in terms of *length* or *arclength*, regions in the plane or surfaces in $\mathbf{R}^2$ are measured in terms of *area* and regions in $\mathbb{R}^3$ are measured in terms of **volumes**. In other words, the **measure** or size of a set S of points is determined by its so-called **topological dimension**: curves have dimension one, areas have dimension two and volumes have dimension three. (Note: Points have topological dimension zero, hence zero length.)

The "nicest" sets to deal with in $\mathbb{R}^k$ are rectangular sets, defined below for $k = 1$ and 2:

1. In $\mathbb{R}$: The interval $[a, b]$, i.e., $a \leq x \leq b$, with length $L = b - a$.

2. In $\mathbb{R}^2$: The rectangular region $S = \{(x, y) \in \mathbb{R}^2 \mid a_1 \leq x \leq b_1, a_2 \leq x \leq b_2\}$ with area $A = (b_1 - a_1)(b_2 - a_2)$.

In general, for $\mathbb{R}^k$, $k = 1, 2, 3, \ldots$, let $S = \{(x_1, x_2, \ldots, x_k) \,|\, x_i \in [a_i, b_i], 1 \leq i \leq k\}$. Then the k-dimensional "size" of $S \in \mathbb{R}^k$, which we shall denote as $L^k(S)$, is given by

$$L^k(S) = \prod_{i=1}^{k} (b_i - a_i). \tag{66}$$

The "L" is in honour of the mathematician H. Lebesgue who was one of the pioneers in the mathematical theory of integration and measure. The quantity $L^k(S)$ is referred to as the "k-dimensional Lebesgue measure of S."

We now consider arbitrary sets $S \in \mathbb{R}^k$ and the problem of determining their sizes. Of course, to measure a set S in $\mathbb{R}^k$, one must use an appropriate k-dimensional "measuring device:"

1. "measuring rods" for lengths ($k = 1$),

2. "measuring tiles" for areas ($k = 2$),

3. "measuring blocks" for volumes ($k = 3$).

Indeed, you may recall that the use of infinitesimal measuring devices, e.g.,

1. dx in $\mathbb{R}$,

2. $dA = dx\, dy$ in $\mathbb{R}^2$,

3. $dV = dx\, dy\, dz$ in $\mathbb{R}^3$,

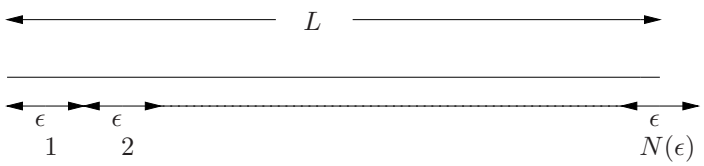
is an example of the "spirit of Calculus" in determining lengths, areas and volumes, not to mention general integrals of scalar and vector functions over regions in $\mathbb{R}^k$.

As a very simple example of this measurement procedure, suppose we wish to determine the length L of a line, e.g. an interval S on $\mathbb{R}$. Now assume that we have a measuring rod of length ϵ, where ϵ is small compared to L. We count the number $N(\epsilon)$ of these rods necessary to "cover" the interval, i.e. beginning at one end of the line, we place the rods next to each other (or slide the rod one length ϵ toward the other end) until we reach the other end, as shown in the figure below. Of course, it is most probable that only a fraction of the final rod is necessary to arrive at the endpoint, i.e. the number of rods $N(\epsilon)$ needed to cover S is not an integer. We shall simply assume that ϵ is sufficiently small to make $N(\epsilon)$ so large that a roundoff to the nearest integer is negligible. (In any case, we shall be concerned with the limit $\epsilon \to 0$ in a later discussion.) Of course, our estimate of the length of S is

$$L(\epsilon) = N(\epsilon)\epsilon. \tag{67}$$

We also expect that

$$\lim_{\epsilon \to 0} L(\epsilon) = L. \tag{68}$$



Now suppose that we shrink our measuring rod by a factor $0 < r < 1$ to produce a shorter rod of length $\epsilon' = r\epsilon$. The question is, "What is $N(\epsilon')$, the number of rods of length ϵ' needed to cover I?" The answer is

$$N(\epsilon') = N(r\epsilon) = N(\epsilon) r^{-1}. \tag{69}$$

This result is "intuitively obvious": If you shrink your measuring rod by a factor of $r = \frac{1}{2}$, then you'll clearly need twice the number of rods of this length. But here is another way to look at it. Suppose that ϵ is sufficiently small that the estimates of length yielded by both the ϵ as well as $\epsilon' = r\epsilon$ measuring rods are, for all intents and purposes, considered to be equal to the actual length L. Then,

$$N(\epsilon)\epsilon = N(r\epsilon)(r\epsilon) = L\,. \tag{70}$$

If we divide both sides of the first equation by $r\epsilon$, we obtain Eq. (69).

Now consider this method of measurement in $\mathbb{R}^2$. Let S denote a region with unknown area A. Suppose that we use square tiles with sides of length ϵ, hence area ϵ^2. (Again assume that ϵ is small.) Let $N(\epsilon)$ be the number of square tiles needed to "cover" the set S. (This is equivalent to drawing region S on graph paper with grid spacing ϵ and counting the number of squares that contain points in S.) Then the estimate of the area A is

$$A(\epsilon) = N(\epsilon)\epsilon^2. \tag{71}$$

We also expect that

$$\lim_{\epsilon \to 0} A(\epsilon) = A. \tag{72}$$

Now suppose that we once again shrink our measuring apparatus. In this case, we shrink *both sides* of our square tile by a factor $0 < r < 1$ to produce a smaller tile with sides of length $\epsilon' = r\epsilon$, hence area $(\epsilon')^2 = r^2\epsilon^2$. Once again, the question is, "What is $N(\epsilon')$, the number of rods of length ϵ' needed to cover I?" The answer is

$$N(\epsilon') = N(r\epsilon) = N(\epsilon) r^{-2}. \tag{73}$$

We can derive this result in the same way as was done for Eq. (69) above. Assume that ϵ is sufficiently small that

$$N(\epsilon)\epsilon^2 = N(r\epsilon)(r\epsilon)^2 = A\,. \tag{74}$$

Division of both sides of the first equation by $(r\epsilon)^2$ yields Eq. (73).

The generalization to "nice" sets in $\mathbb{R}^k$, $k = 1, 2, 3, \ldots$, is quite straightforward. We work with k-dimensional cubes whose sides have length ϵ. hence k-dimensional measure or volume ϵ^k. Suppose that $N(\epsilon)$ of these cubes are needed to cover a set $S \in \mathbb{R}^k$. Then the estimate of the k-dimensional volume of S is

$$L^k(S) \approx N(\epsilon)\epsilon^k \tag{75}$$

and we expect that

$$\lim_{\epsilon \to 0} N(\epsilon)\epsilon^k = L^k(S). \tag{76}$$

If we shrink **all** sides of our k-dimensional (hyper)cube by the factor $0 < r < 1$ to produce a smaller (hyper)cube with sides of length $\epsilon' = r\epsilon$ and k-dimensional volume $(\epsilon')^k = r^k\epsilon^k$, then the number $N(\epsilon')$ of these smaller cubes needed to cover S is given by

$$\boxed{N(\epsilon') = N(r\epsilon) = N(\epsilon)r^{-k}.} \tag{77}$$

Because we have grown up with the notions of length, area and volume of "nice," homogeneous sets, the result summarized in Eq. (77) is viewed as obvious. However, we now turn to the construction of sets in $\mathbb{R}^k$ for which this relation does not necessarily apply. It turns out that we shall encounter difficulties because our "nice" Euclidean measuring sets, i.e. rods of length ϵ, squares of area ϵ^2, cubes of volume ϵ^3, etc., may be adequate to cover a set S but inadequate to assign a nonzero or noninfinite volume to it in the limit $\epsilon \to 0$.

Does this already sound unpalatable? It shouldn't, since such problems are also encountered in the Euclidean domain:

1. **Question:** What is the area of a line segment that lies in the plane $\mathbb{R}^2$? **Answer:** Zero.

2. **Question:** What is the length of a triangular region that lies in the plane? **Answer:** Infinity.

In other words, in order to obtain both a finite and nonzero result, we have to fit the measuring apparatus to the "topological nature" of the set we are measuring: We measure lengths with rods, areas with squares, volumes with cubes.

In closing, consider the following example. Let $S = \{(x, y) \mid 0 \leq x, y \leq 1\}$, the unit square in $\mathbb{R}^2$. It should not be too difficult to see that

- The length of S (its one-dimensional, or L^1, size) is infinite.
- The area of S (its two-dimensional, or L^2, size) is 1.
- The volume of S (its three-dimensional, or L^3, size) is 0.

Mathematically, we may summarize these results as follows,

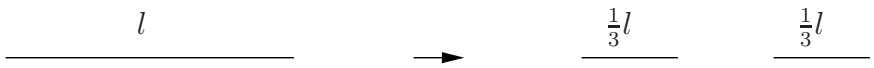
$$L^1(S) = \infty, \qquad L^2(S) = 1, \qquad L^3(S) = 0.$$

This simple example illustrates that in order to obtain a "acceptable" result for the size of a set, i.e., one that is neither zero nor infinite, but rather a finite number, one must use an appropriate "measuring apparatus", i.e., an apparatus of **appropriate dimension**, namely, a dimension which matches the dimensionality of the set S one is measuring. In the case of "nice," Euclidean sets, the dimension is an integer. But in the case of "non-Euclidean" or "irregular" sets, the dimension is not necessarily integer-valued, in which case we consider S to be a "fractal" set, i.e., a set of **fractional dimension**.

Lecture 26

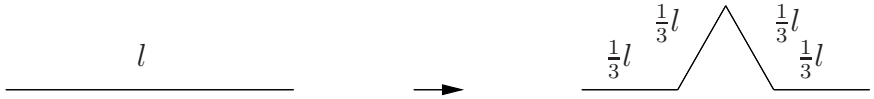
Measuring the sizes of "fractal" sets in $\mathbb{R}^k$

First recall the construction of the ternary Cantor set $C \subset [0,1]$ discussed in Section 3.7. Recall that C was produced by the repeated application of a "middle-thirds dissection" procedure, shown on Pages 3.62 and 3.63: From $J_0 = [0,1]$, we remove the open set $(\frac{1}{3}, \frac{2}{3})$ to produce $J_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$. From J_1 we remove the open sets $(\frac{1}{9}, \frac{2}{9})$ and $(\frac{7}{9}, \frac{8}{9})$ to produce $J_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$. We may consider this process as the repeated application of a "generator" G that operates on sets of points. Think of the generator as a kind of computer equipped with a pattern recognition device. It searches an input object for line segments. When it finds a line segment of length l, it removes the middle thirds of the line segment. In summary the generator G may be represented by the procedure

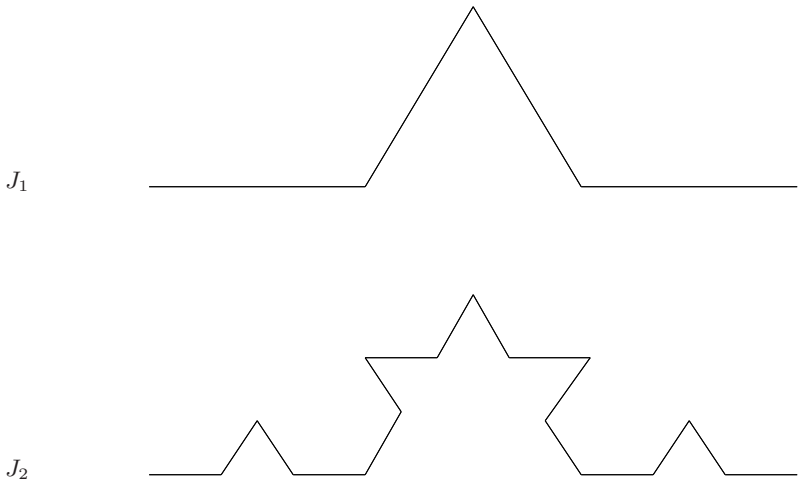


The middle thirds dissection procedure associated with the Cantor set may then be considered as a dynamical system involving the generator G: We use the input seed $J_0 = [0,1]$ and then define the iteration sequence $J_{n+1} = G(J_n)$, $n = 0, 1, 2, \ldots$. In the limit, the sets J_n converge to the Cantor set C.

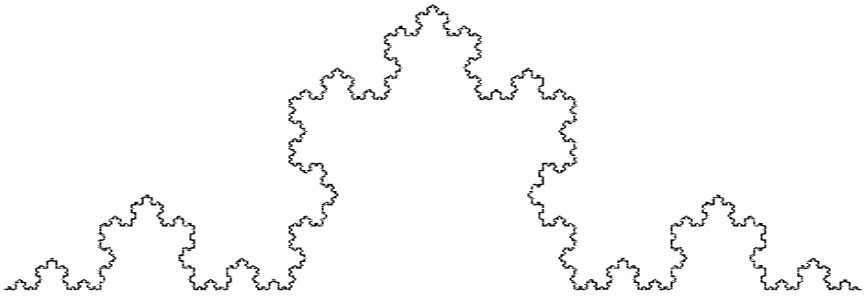
Now consider the following generator that produces a limiting curve in $\mathbb{R}^2$:



This generator removes the middle third of a line segment of length l and replaces it with two segments of length $\frac{l}{3}$. If we begin with the seed $J_0 = [0,1]$, the next two sets in the iteration sequence $J_{n+1} = G(J_n)$ are sketched below.



We state, without proof, that the sets J_n converge to a very complicated curve in $\mathbb{R}^2$, the so-called "von Koch curve" shown below.



The von Koch curve

This curve C is a **fractal curve**. It is a **self-similar set** in the sense that C is a union of four contracted copies of itself, C_i, $1 \leq i \leq 4$. Each copy C_i can be obtained by contracting C, say toward its left-most point (here considered as the origin of our coordinate system), and then translating and/or rotating the contracted copy appropriately. It is not surprising that the positions of the four copies C_i are determined by the four components of the piecewise linear curve $J_1 = G(J_0)$ shown earlier.

The copying property of the curve C can be continued: Since each subset C_i may be expressed as a union of four contracted copies C_{ij} of itself, the original curve C may be expressed as a union of sixteen contracted copies of itself.

Let us now try to measure the length of the von Koch curve, using the method of ϵ-measuring rods discussed in the previous section. We start at the left endpoint of C and place these rods end to end until we reach the right endpoint of C. A most natural choice for the lengths of these rods, with increasing refinement, i.e. $\epsilon \to 0$, is the sequence $\epsilon_n = (\frac{1}{3})^n$, with $n \to \infty$.

If we use a measuring rod of length $\epsilon_0 = 1$, then clearly only one such rod is needed to get us from the left endpoint of C to the right endpoint. Clearly, this length estimate, $L_0 = 1$, is a very gross one since it ignores the quite significant middle "inlet" of the curve, not to mention all the tinier "nooks and crannies" that are present in the curve at all sizes. If we use a measuring rod of length $\epsilon_1 = \frac{1}{3}$, then four rods are necessary to get us from the left endpoint of C to the right endpoint, yielding a length estimate of $L_1 = \frac{4}{3}$. This rod at least detects the middle "inlet." Nevertheless, it cannot detect any smaller inlets. It should be clear that $4^2 = 16$ measuring rods of length $\epsilon_2 = (\frac{1}{3})^2$ are necessary to cover C, yielding a length estimate of $L_2 = \frac{16}{9} = (\frac{4}{3})^2$.

The reader should notice that the "covering" of C with ϵ_n-measuring rods yields length estimates

L_n that are equivalent to the lengths of the curves J_n obtained by the repeated action of the generator G above. The results are summarized below:

Length of measuring rod: $\epsilon_n = \frac{1}{3^n}$
Number of measuring rods needed to cover C: $N(\epsilon_n) = 4^n$
Estimate of length: $L_n = N(\epsilon_n)\epsilon_n = (\frac{4}{3})^n$.

Therefore, we conclude that the length of C is given by

$$L^1(C) = \lim_{n\to\infty} L_n = \infty, \tag{78}$$

i.e. the von Koch curve has infinite length. In retrospect, we are not surprised since the generator essentially increases the length of each curve J_k by a factor of $\frac{4}{3}$ to produce the curve J_{k+1}. However, the fact that C has infinite length is quite intriguing. In fact, the results summarized above already indicate that something "non-Euclidean" is going on here. Note the relationship between consecutive numbers of ϵ_n-rods necessary to cover C:

$$N(\epsilon_{n+1}) = 4N(\epsilon_n). \tag{79}$$

For any n, we may set $\epsilon = \epsilon_n$ and $\epsilon' = r\epsilon$, where $r = \frac{1}{3}$, to obtain the relation

$$N\left(\frac{\epsilon}{3}\right) = 4N(\epsilon), \tag{80}$$

which does **not** agree with Eq. (69) for one-dimensional objects, since the factor "4" would have to be replaced by "3." Once again, we suspect that something "non-Euclidean" is going on here.

One may well wonder whether these deviations are due to the fact that the "curve" C may have some area, due to its complicated jaggedness. In fact, Eq. (79) indicates that this is not the case, since, from Eq. (73), the factor "4" would have to be replaced by "9." Nevertheless, we attempt to estimate the area of C using the method of ϵ-tiles of area ϵ^2 outlined earlier. The reader may worry that the covering of C with square tiles of size $\epsilon_n \times \epsilon_n$ will yield gross overestimates of the area, since the tiles overlap significantly with each other. (One can choose to cover C, not with squares but with **circular discs** of diameter ϵ_n. The resulting estimate will differ only by a factor of $\frac{\pi}{4}$. Why?) Our estimate will be, up to a constant factor,

$$L^2(C) \cong N(\epsilon_n)\epsilon_n^2. \tag{81}$$

A little reflection will reveal that the number of ϵ_n-sided square tiles or ϵ_n-diameter circular discs required to cover C is identical to the number $N(\epsilon_n)$ of ϵ_n-rods needed to "cover" C when we were approximating its area. Therefore, we have the following results:

Area of disc (up to factor K): $\epsilon_n^2 = (\frac{1}{3^n})^2$

Number of discs needed to cover C: $N(\epsilon_n) = 4^n$
Estimate of area: $A_n = N(\epsilon_n)\epsilon_n^2 = (\frac{4}{9})^n$.

We conclude that the area of C is given by

$$L^2(C) = \lim_{n\to\infty} L_n = 0. \tag{82}$$

Even an overestimate of the area yields a result of zero. We have found that the von Koch curve C has infinite length yet zero area! An explanation of these two results, in terms of our measuring methods, is rather simple. In both cases, $N(\epsilon_n) = 4^n$ measuring rods or discs are required to cover C at each resolution. We note that

$$\left(\frac{4}{3}\right)^n \to \infty \qquad \text{and} \qquad \left(\frac{4}{3^2}\right)^n \to 0, \qquad \text{as} \qquad n \to \infty. \tag{83}$$

In the first case, the factor "3" is not large enough to match the factor "4" to produce a noninfinite limit. In the second case, the factor "3" is raised to too high a power, i.e. 3^2, to produce a nonzero limit. One may suspect that if we were allowed to raise the factor "3" to a suitable power, say $1 < D < 2$, we could "steer" the above limit away from both zero and infinity. In fact, if we choose D so that

$$3^D = 4\,, \tag{84}$$

then the ratio in Eq. (83) becomes

$$\frac{4}{3^D} = 1 \tag{85}$$

so that

$$\lim_{n\to\infty} \frac{4}{3^D} = \lim_{n\to\infty} 1^n = 1\,. \tag{86}$$

We may easily solve for D in Eq. (84) by taking logarithms of both sides. Any logarithm will do, but we'll take natural logarithms,

$$\ln(3^D) = \ln 4 \quad \Longrightarrow \quad D\,\ln 3 = \ln 4\,, \tag{87}$$

which implies that

$$D = \frac{\ln 4}{\ln 3} \approx 1.26\,. \tag{88}$$

Since the dimension D lies between 1 and 2, we may consider the von Koch curve as being a set which is "thicker" than a curve but "thinner" than a region in the plane.

This is generally the situation encountered with fractal sets S that are "embedded" or constructed in the underlying space $\mathbb{R}^k$, k integer - in this case $k = 2$, i.e. the plane. It is typically found that

$$L^{k-1}(S) = \infty, \qquad L^k(S) = 0. \tag{89}$$

In other words, the set S is thicker than a $(k-1)$-dimensional set, but thinner than a k-dimensional set. Its "dimension" lies somewhere between $k-1$ and k.

This motivated mathematicians to consider the following "q-dimensional measure" of a set S, where $q \in \mathbb{R}$ is allowed to be any real number:

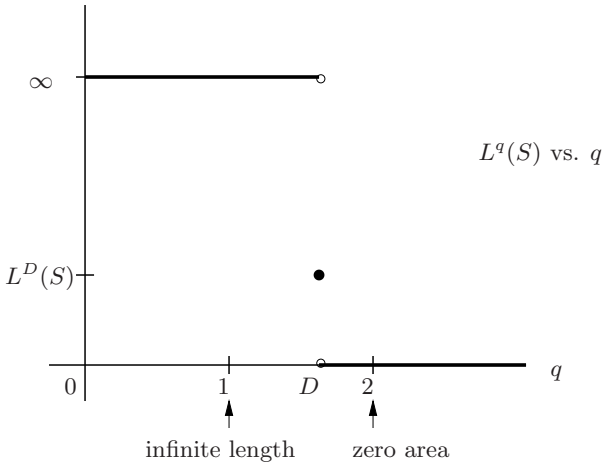
$$L^q(S) = \lim_{\epsilon \to 0} N(\epsilon)\, \epsilon^q \,. \tag{90}$$

Here, $N(\epsilon)$ is the number of lines/tiles/cubes/whatever needed to cover the set S.

We now state, without proof, that for "self-similar" fractal sets $S \in \mathbb{R}^k$ such as the Cantor set and the von Koch curve the following holds: There exists a value D, $k-1 \leq D \leq k$, such that

$$L^q(S) = \infty, \quad q < D, \qquad L^q(S) = 0 \quad q > D, \qquad q \in \mathbb{R}. \tag{91}$$

The value D is called the "fractal dimension" (or "capacity dimension") of the fractal set S. This generic behaviour of the q-dimensional measure function L^q is sketched below.



To summarize: **The fractal set S is "thicker" than any set of dimension $q < D$ and "thinner" than any set of dimension $q > D$.**

Now recall the example given at the end of the previous section, in which we considered a subset S of the plane that had finite area. The structure of S - a simple planar, "Euclidean," region - was such that the dimension $D = 2$ was appropriate to describe its "size." We could work with integer values to zero in on its dimension, determining that its length ($q = 1$) was infinite and its volume ($q = 3$) was zero. However, integer values of dimension are not enough to analyze fractal sets such as the von Koch curve - we must go to non-integer values to zero in on an appropriate "dimension" of the set.

For most self-similar fractal sets S, their appropriate "size" is then given by the D-dimensional volume or measure

$$L^D(S) = \lim_{\epsilon \to 0} N(\epsilon)\epsilon^D. \tag{92}$$

Let's now see the consequence of this result. For ϵ sufficiently close to zero, the D-dimensional volume or measure of a set S is well approximated as follows,

$$L^D(S) = N(\epsilon)\epsilon^D = N(r\epsilon)(r\epsilon)^D \,. \tag{93}$$

If we divide both sides by $(r\epsilon)^D$, we obtain the following important result,

$$\boxed{N(r\epsilon) = N(\epsilon)r^{-D}\,,} \tag{94}$$

which is a scaling relationship between the number of ϵ-tiles and the number of $r\epsilon$-tiles needed to cover set S. Note that this is a non-integer generalization of the scaling relationship (77) observed for Euclidean sets, i.e.,

$$N(r\epsilon) = N(\epsilon)r^{-k}\,, \tag{95}$$

for k an integer.

Return to von Koch curve: In this case $r = \frac{1}{3}$. If $N(\epsilon)$ is the number of ϵ-tiles needed to cover the von Koch curve C, then the number of $N(r\epsilon) = N(\frac{1}{3}\epsilon)$ tiles needed to cover C is $4N(\epsilon)$. Eq. (94) then becomes

$$N\left(\frac{1}{3}\epsilon\right) = 4N(\epsilon) = N(\epsilon)\left(\frac{1}{3}\right)^{-D} \,. \tag{96}$$

Division of both sides of the second equality by $N(\epsilon)$ and a little rearrangement yields

$$3^D = 4 \quad \Longrightarrow \quad D = \frac{\ln 4}{\ln 3}\,, \tag{97}$$

which is in agreement with the value of D in Eq. (88).

Return to ternary Cantor set C on [0,1]: In Section 3.7, we found that its length was zero. Recall that this followed from a look at the lengths of the sets J_n that converge to C in the limit $n \to \infty$: $L^1(J_n) = \frac{2}{3}^n$. In fact these are precisely the length estimates of the Cantor set that would be obtained if we used ϵ measuring rods of length $\epsilon_n = \frac{1}{3^n}$. In summary:

Length of measuring rod: $\epsilon_n = \frac{1}{3^n}$
Number of measuring rods needed to cover C: $N(\epsilon_n) = 2^n$
Estimate of length: $L_n = N(\epsilon_n)\epsilon_n = (\frac{2}{3})^n$.

It follows that the length of C is given by

$$L^1(C) = \lim_{n\to\infty} L_n = 0. \tag{98}$$

It now remains to determine the fractal dimension of the Cantor set C. Once again, we begin with the fact that the number of square tiles of length $\epsilon_n = \frac{1}{3^n}$ needed to cover C is the same as the number of rods we used above: $N(\epsilon_n) = 2^n$. We may now proceed in either of two ways (each of which, of course, is related to the other):

1. From $N(\epsilon_n) = 2^n$, it follows that

$$\begin{aligned} N(\epsilon_{n+1}) &= N(\frac{1}{3}\epsilon_n) \\ &= 2N(\epsilon_n) \end{aligned} \tag{99}$$

Insert these results into Eq. (94):

$$N\left(\frac{1}{3}\epsilon\right) = 2N(\epsilon) = N(\epsilon)\left(\frac{1}{3}\right)^{-D} . \tag{100}$$

Division of both sides of the second equality by $N(\epsilon)$ and a little rearrangement yields

$$3^D = 2 \quad \Longrightarrow \quad D = \frac{\ln 2}{\ln 3} \approx 0.63 . \tag{101}$$

2. Since the q-dimensional measure of C, $q \in \mathbb{R}$, will be given by

$$\begin{aligned} L^q(C) &= \lim_{n\to\infty} N(\epsilon_n)(\epsilon_n)^q \\ &= \lim_{n\to\infty} \left(\frac{2}{3^q}\right)^n , \end{aligned} \tag{102}$$

it follows that the critical value for q occurs when $3^q = 2$ or $q = \frac{\ln 2}{\ln 3}$. When $q < \frac{\ln 2}{\ln 3}$, $L_q(C) = \infty$. When $q > \frac{\ln 2}{\ln 3}$, $L_q(C) = 0$. Therefore, $D = \frac{\ln 2}{\ln 3}$, in agreement with the first method.

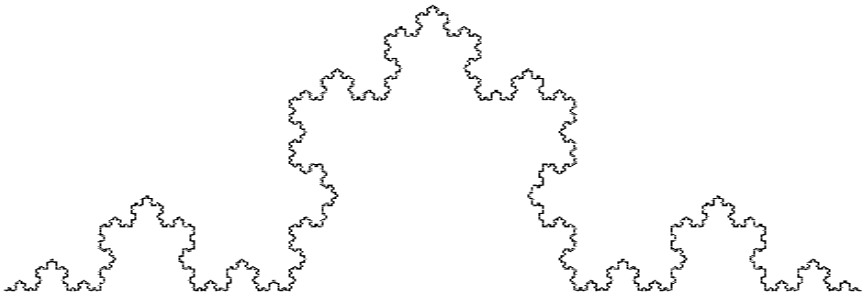
The fact that the fractal dimension D of the Cantor set lies between 0 and 1 suggests that it is "thicker" than a set of points but "thinner" than a curve/line.

A summary of the above results

A summary is in order, in an effort to get an idea of the "big picture."

- When the following generator, is applied repeatedly to the unit interval $J_0 = [0, 1]$ the resulting

  sequence of sets J_n defined by the iteration sequence $J_{n+1} = G(J_n)$, $n = 0, 1, 2, \cdots$, converges to the following set,



  known as the **von Koch curve**. The length of the von Koch curve was found to be infinite, and its area was found to be zero. Using a scaling argument, we found its fractal dimension to be

$$D = \frac{\ln 4}{\ln 3} \approx 1.26\,. \tag{103}$$

  Since $D > 1$, this set may be viewed as "thicker" than a curve but "thinner" than a region in the plane.

  Finally, the von Koch curve may be viewed as **a union of four copies of itself**, each copy of which is obtained by shrinking the von Koch curve by a factor $r = \frac{1}{3}$ in each direction.

- When the following generator,

  is applied repeatedly to the unit interval $J_0 = [0, 1]$ the resulting sequence of sets J_n defined by the iteration sequence $J_{n+1} = G(J_n)$, $n = 0, 1, 2, \cdots$, converges to the following set,

known as the **ternary Cantor set**. The length of the Cantor set was found to be zero. Using a scaling argument, we found its fractal dimension to be

$$D = \frac{\ln 2}{\ln 3} \approx 0.63\,. \tag{104}$$

Since $D < 1$, this set may be viewed as "thinner" than a curve, which has topological dimension one. Furthermore, since $D > 0$, this set may be viewed as "thicker" than a set of points which have topological dimension zero.

Finally, the von Koch curve may be viewed as **a union of two copies of itself**, each copy of which is obtained by shrinking the von Koch curve by a factor $r = \frac{1}{3}$ in each direction.

Let us now consider the following generator, which may be viewed as a kind of "middle road" between the two generators above – a generator which neither adds nor removes sets from a set:

Yes, this may be viewed as a generator: It simply returns the "middle-thirds" part of an interval that the Cantor set generator removed. In effect, it represents the identity operator. It simply maps the unit interval $J_0 = [0, 1]$ to itself. In the limit, the sequence of sets $J_{n+1} = G(J_n)$ converges trivially to the interval $[0, 1]$. Of course, the dimension of this set is $D = 1$. But let's derive this result using the scaling associated with this generator. (We actually did this in the previous lecture.)

We view the unit interval $[0, 1]$ as **a union of three copies of itself**, each copy of which is obtained by shrinking the interval by a factor of $r = \frac{1}{3}$. Let $N(\epsilon)$ be the number of ϵ-sticks needed to cover the unit interval $[0, 1]$. If we now use measuring sticks of length $\frac{1}{3}$, we need $3N(\epsilon)$ sticks to cover the interval, i.e.,

$$N\left(\frac{1}{3}\epsilon\right) = 3N(\epsilon)\,. \tag{105}$$

We now use the scaling result derived earlier,

$$N(r\epsilon) = N(\epsilon) r^{-D}\,, \tag{106}$$

with $r = \frac{1}{3}$ in Eq. (105):

$$N\left(\frac{1}{3}\epsilon\right) = 3N(\epsilon) = N(\epsilon)\left(\frac{1}{3}\right)^{D}\,. \tag{107}$$

Dividing both sides of the final equation by $N(\epsilon)$ and rearranging yields

$$3^D = 3 \quad \Longrightarrow \quad D = 1\,. \tag{108}$$