1992

An Investigation of Iterated Function Systems and Fractals

David Wuolu
College of Saint Benedict/Saint John's University, dwuolu@csbsju.edu

AN INVESTIGATION OF ITERATED FUNCTION SYSTEMS AND FRACTALS

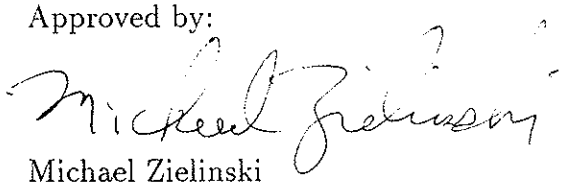
A THESIS
The Honors Program
College of St. Benedict/St. John's University

In Partial Fulfillment
of the Requirements for the Distinction "All College Honors"
and the Degree Bachelor of Arts
In the Department of Mathematics

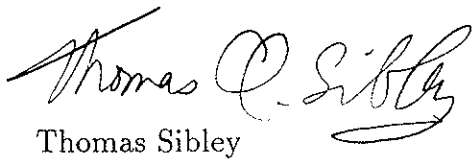
by
David J. Wuolu
April, 1992

Project Title: An Investigation of Iterated Function Systems and Fractals

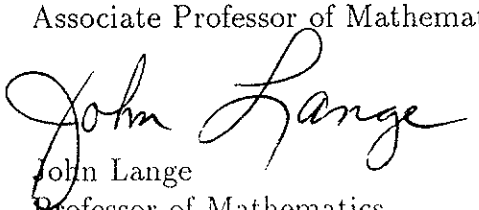
Approved by:



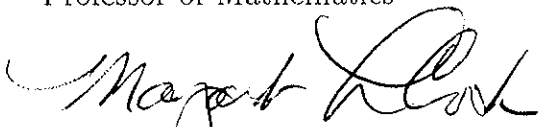
Michael Zielinski
Assistant Professor of Mathematics



Thomas Sibley
Associate Professor of Mathematics (Chair)



John Lange
Professor of Mathematics



Margaret Cook
Director, Honors Program

Acknowledgements

This essay is the result of research under Michael F. Zielinski, to whom I am doubly indebted - both for the time and patience with which he explained difficult concepts to a less than patient student, and for his effort in writing computer programs which enabled us to visualize the wonderful process of fractal creation.

I would like to also thank Michael Gass, Thomas Sibley, and John Lange for reading and correcting my many errors as well as teaching me how to write.

I am also indebted to Lorie J. Warren, my fiance, for her infinite patience and encouragement throughout the long and sometimes arduous process of writing this paper.

Finally, I would like to thank my parents, Bob and Anne, for making the sacrifices they did to give me the education I have received.

This paper explores some of the fascinating ideas involved in the field of dynamical systems and fractals. Exciting new discoveries in the past fifteen years have opened up an entire new world of research opportunities. The beautiful computer generated pictures of fractals enable mathematicians to see the intrinsic beauty of each system on their own desktops.

The history behind fractals and dynamical systems is fragmented, but the seeds from which the modern theory stems were planted in the previous century by Poincare and in the early twentieth century by Fatou and Julia. The purpose of this paper is not to be a historical account of fractal evolution, however, and we will get directly to the mathematics itself. I have hoped to make the vast majority of my paper accessible to those with a strong calculus and linear algebra background. There are some proofs which require analysis, but by carefully reading what the theorem says, one is able to understand the material and gain insight into the process without the proof.

I have drawn primarily from the work of Michael Barnsley, who adopts a framework of metric spaces from which to approach the ideas of fractal geometry. First it is good to start with a quick overview of the investigation. Concepts which are important include metric spaces, open and closed sets, compactness, convergence, connectedness, completeness and Cauchy sequences, to name a few. A metric space of particular importance is $\mathcal{H}(\mathbf{X})$, which is the set of all compact subsets of a metric space.

Once these concepts are understood, we move on to transformations on metric spaces. We will define fractals using iterated function systems (IFS's), and many interesting theorems are proven.

Definition [Kirkwood, p. 23]: *A metric space* ($\mathbf{X}$,d) is a space $\mathbf{X}$ together with a real-valued function $d : \mathbf{X} \times \mathbf{X} \rightarrow \boldsymbol{R}$, which defines the distance between pairs of points x and y in $\mathbf{X}$. We require that d obeys the following axioms:

1. $d(x,y) = d(y,x) \ \forall \ x,y \in \mathbf{X}$
2. $0 < d(x,y) < \infty \ \forall \ x,y \in \mathbf{X}, x \neq y$
3. $d(x,x) = 0 \ \forall \ x \in \mathbf{X}$
4. $d(x,y) \leq d(x,z) + d(z,y) \ \forall \ x,y,z \in \mathbf{X}$.

Such a function d is called a *metric*. Well known examples of metric spaces include ($\boldsymbol{R}$,Euclidean) and ($\boldsymbol{R}$,Manhattan). The Euclidean metric is:

$$d(x,y) = \mid x - y \mid \text{ for } \boldsymbol{R} \text{ and } d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad \text{for } \boldsymbol{R}^2.$$

To see that the Euclidean metric on $\boldsymbol{R}^2$ is indeed a metric, note that for $x = (x_1, x_2)$ and $y = (y_1, y_2)$:

1. $d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2} = d(y,x)$.

2. $0 < d(x,y)\ \forall x,y\ \in \mathbf{X}, x \neq y$ since the square root of the sum of squares is always non-negative. $d(x,y) < \infty\ \forall\ x,y \in \mathbf{X}\ ,\ x \neq y$ by inspection.

3. $d(x,x) = \sqrt{(x_1 - x_1)^2 + (x_2 - x_2)^2} = 0, \forall\ x \in \mathbf{X}$.

4. The proof of this involves the triangle inequality and can be found in [Kirkwood, p. 22], and many other mathematics texts. q.e.d.

The Manhattan (or Taxicab) metric for $\boldsymbol{R}$ is the same as the Euclidean, but for $\boldsymbol{R}^2$ becomes:

$$d(x,y) = \mid x_1 - y_1 \mid + \mid x_2 - y_2 \mid .$$

One distance function in $\boldsymbol{R}$ which is not a metric is:

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

This fails the fourth requirement as the following is a counterexample:

$$d(-2,3) = 6 \not\leq 0 = d(-2,0) + d(0,3).$$

Here we will introduce a very important space, Σ_N, called code space on N symbols, where N is a positive integer.

Definition [Barnsley, p. 10]: *Code Space* on N symbols, Σ_N, is defined to be the space of all possible infinitely long strings of numbers made from the symbols 0,1,2,...,N-1.

Hence, an example of a point in code space on 3 symbols is:

$$\text{x} = 10012001201212012002102020220121122 0...$$

In general, we write

$$\text{x} = x_1x_2x_3x_4... \text{ where each } x_i \in 0,1,2,...,N-1.$$

Definition [Barnsley, p. 12]: The *Code Space Metric*, d_c, is defined on Σ_N to be:

$$d_c(x,y) = d_c(x_1x_2x_3..., y_1y_2y_3...) = \sum_{i=1}^{\infty} \frac{|x_i - y_i|}{(N+1)^i}.$$

Theorem: Every two points in Σ_N are a finite distance apart. In fact, all points are within one of each other.

To see this note that:

$$\text{Max}\{d_c(x,y)\} = \sum_{i=1}^{\infty} \frac{N-1}{(N+1)^i} = \frac{N-1}{N+1}(1 + \frac{1}{N+1} + \frac{1}{(N+1)^2} + \cdots) =$$

$$\frac{N-1}{N+1}(\frac{1}{1-\frac{1}{N+1}}) = \frac{N-1}{N+1}(\frac{N+1}{N}) = \frac{N-1}{N} < 1. \text{ q.e.d.}$$

Another useful result of the code space metric is that it makes it very easy to know what points near each other look like. Suppose we are working with Σ_3 and want to know what all the point within ϵ of a given point x look like.

Theorem: Let ϵ be of the form $\frac{1}{4^{n+1}}$, x and $y \in \Sigma_3$. All the points y that are within ϵ of x have the first n terms in common with x. Also, if x and y have the first $n+1$ terms in common with each other, then they are within ϵ

For the proof, we take a contrapositive approach, and suppose that for some $j \in \{1, 2, \cdots, n\}$ we have $x_j \neq y_j$ Then, clearly, if $x_j \neq y_j$, $d_c(x,y) = \sum_{n=1}^{\infty} \frac{|x_n - y_n|}{4^n} \geq \frac{1}{4^j} > \frac{1}{4^{n+1}}$. This is true for j up to n, for suppose $x_n \neq y_n$. Then $\sum_{n=1}^{\infty} \frac{|x_n - y_n|}{4^n} \geq \frac{1}{4^n} > \frac{1}{4^{n+1}}$. Hence,

$$d_c(x,y) < \frac{1}{4^{n+1}} \Rightarrow \{x_1 = y_1, \cdots, x_n = y_n\}.$$

Also, if $x_1 = y_1, x_2 = y_2, \cdots, x_{n+1} = y_{n+1}$ then we know that

$$d_c(x,y) < \sum_{i=n+2}^{\infty} \frac{2}{4^i} = \frac{2}{4^{n+2}}(1 + \frac{1}{4} + \frac{1}{4^2} + \cdots) = \frac{2}{3 \cdot 4^{n+1}} < \frac{1}{4^{n+1}}. \quad \text{q.e.d.}$$

Here are some important definitions we shall require in our study of metric spaces:

Definition [Barnsley, p. 19]: Let $S \subset \mathbf{X}$ be a subset of a metric space $(\mathbf{X}, d)$. A point $x \in \mathbf{X}$ is called a *limit point* of S if there is a sequence $\{x_n\}_{n=1}^{\infty}$ of points $x_n \in \{S - x\}$ such that $\lim_{n \to \infty} x_n = x$.

Definition [Kirkwood, p. 60]: S is *open* if for each $x \in S$ there is an $\epsilon > 0$ such that the ball
$B(x, \epsilon) = \{y \in \mathbf{X} : d(x,y) < \epsilon\} \subset S$.

Definition [Barnsley, p. 19]: S is *closed* if it contains all of its limit points.

Definition [Kirkwood, p. 46]: A sequence $\{x_n\}_{n=1}^{\infty}$ of points in a metric space $(\mathbf{X}, d)$ is a *Cauchy sequence* if, for any $\epsilon > 0$, there is an integer $N > 0$ such that

$$d(x_n, x_m) < \epsilon \text{ for all } n, m > N.$$

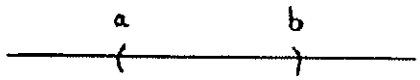
Definition [Barnsley, p. 18]: A metric space $(\mathbf{X}, d)$ is *complete* if every Cauchy sequence $\{x_n\}_{n=1}^{\infty}$ in $\mathbf{X}$ has a limit $x \in \mathbf{X}$.

It is of interest to note that if $(\mathbf{X}, d)$ is a metric space then $\mathbf{X}$ is open and closed. $\mathbf{X}$ is open since any ϵ-ball centered at a point $x \in \mathbf{X}$ is also in $\mathbf{X}$. It is closed because it not only contains all its limit points, it contains "all" points. There are metric spaces which are not complete. For example, take the space defined by $\{\frac{1}{n}\}_{n=0}^{\infty}$ excluding $\{0\}$. This subset of the reals does not contain a limit point of the Cauchy sequence

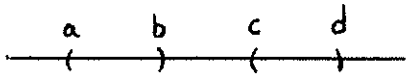
$\{\frac{1}{x}\}_{x=1}^{\infty}$. Hence, it is not complete. Another familiar example of an incomplete metric space is the subset of rationals.

Definition [Kirkwood, p. 69]: A metric space $(\mathbf{X}, d)$ is *connected* if the only two subsets of $\mathbf{X}$ that are simultaneously open and closed are $\mathbf{X}$ and $\emptyset$. A subset $S \subset \mathbf{X}$ is connected if (S, d) is a connected metric space. It is *disconnected* if it is not connected.

An example of a connected metric space is an interval of the real line, say (a,b). An example of a disconnected metric space is the union of disjoint intervals of the real line, say (a,b) ∪ (c,d).



connected



disconnected

Sometimes a metric space is in so many "pieces" that we say it is totally disconnected.

Definition [Barnsley, p. 25]: S is *totally disconnected* if the only nonempty connected subsets of S are subsets consisting of single points.

Theorem [Wuolu]: Code space on N symbols is totally disconnected.

We will do the proof for Σ_3, from which it can easily be generalized to Σ_N. We will need to define the notion of an open ball.

Definition [Wuolu]: We define $B(x, \epsilon)$, an open ball of radius epsilon in some space $\mathbf{X}$, to be
$\{y \in X \mid d(x, y) < \epsilon\}$.

Suppose there existed a nonempty connected subset of Σ_3 which did not consist of a single point. We will denote this set as Ω. We will show that this is disconnected. To do it, we look at the intersection of a ball of radius $\epsilon > 0$ and our subset Ω. Given an $x \in \Omega$ and an $\epsilon > 0$, this ball is denoted by $B_\Omega(x, \epsilon) = \{y \in \Omega \mid d(x, y) < \epsilon\}$. Suppose ϵ is $\frac{1}{4^{j+1}}$, j an integer > 1. Then

$$B_\Omega(x, \Omega) = \{\sigma \in \Sigma_3 \mid \sigma_1 = x_1, \sigma_2 = x_2, \ldots, \sigma_j = x_j\}.$$

To show that Ω is disconnected, our goal, we show that it is not connected, i.e. that there exists a subset which is simultaneously open and closed in Ω besides $\emptyset$ and Ω.

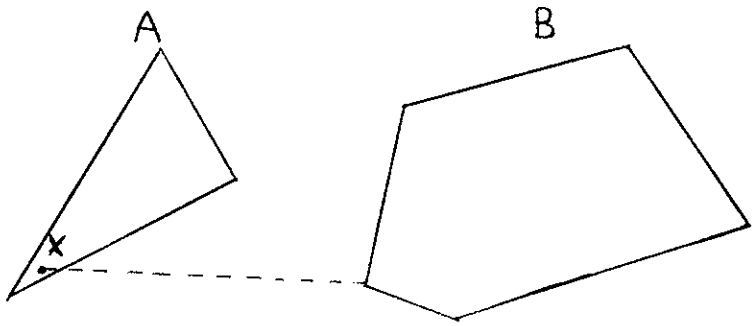
So, pick $\delta = \frac{\epsilon}{4}$. Then $B_\Omega(x,\delta) \subset B_\Omega(x,\epsilon)$. We claim that this δ-ball is both open and closed, and hence Ω is disconnected. To see that $B_\Omega(x,\delta)$ is open, note that $B_\Omega(x,\delta) = \Omega \cap B(x,\delta)$, and by our definition of balls, is open. To see that $B_\Omega(x,\delta)$ is closed, let's show that it contains all of its limit points. Suppose ω is a limit point of $B_\Omega(x,\delta)$. Then, by definition, there is a sequence of points $\{x_n\}_{n=1}^{\infty}$ (where each $x_i \neq \omega$) in $B_\Omega(x,\delta)$ such that $\lim_{n\to\infty} x_n = \omega$. Recall $\delta = \frac{1}{4^{j+2}}$. Then with our metric, if a point y is in $B_\Omega(x,\delta)$, then y has the first $j+1$ terms in common with x. In other words, $y_1 = x_1, \cdots, y_{j+1} = x_{j+1}$. Hence, our sequence of points $\{x_n\}_{n=1}^{\infty}$ which have as their limit point ω, have for each x_n, $x_{n1} = x_1, \cdots, x_{nj+1} = x_{j+1}$. We now pick a very small ϵ_k neighborhood of ω, so that all points in $B_\Omega(\omega,\epsilon_k)$ have the first k terms in common. Since $\{x_n\}_{n=1}^{\infty}$ converges to ω, we can make ϵ_k small enough that $k > j+1$. Hence, if $x_n \in B(\omega,\epsilon_k)$, then $\omega_1 = x_{n1}, \cdots, \omega_j = x_{nj}, \cdots, \omega_k = x_{nk}$. But since x_n has the first $j+1$ terms in common with x and ω has the first $j+1$ terms in common with x_n, ω has the first $j+1$ terms in common with x and hence $\omega \in B_\Omega(x,\delta)$. q.e.d.

When making fractals, we will be dealing with a special type of metric space, called a compact metric space. The reasons for using this type of space will become apparant as we move to an understanding of how fractals are defined.

Definition [Barnsley, p. 20]: Let $S \subset \mathbf{X}$ be a subset of a metric space $(\mathbf{X}, d)$. S is *compact* if every infinite sequence $\{x_n\}_{n=1}^{\infty}$ in S contains a subsequence having a limit in S.

We now have the tools necessary to formally define what is known as the "space of fractals." Let $(\mathbf{X}, d)$ be a complete metric space. Then $\mathcal{H}(\mathbf{X})$ denotes the space whose points are the compact subsets of $\mathbf{X}$, other than the empty set. When dealing with this space, we define the distance from a point x in $\mathbf{X}$ to a point B in $\mathcal{H}(\mathbf{X})$ (which is really a compact set) to be:

$$d(x,B) = min\{d(x,y) : y \in B\}.$$



The compactness and nonemptiness of the set B guarantees that this distance exists, i.e. that there is such a minimum.

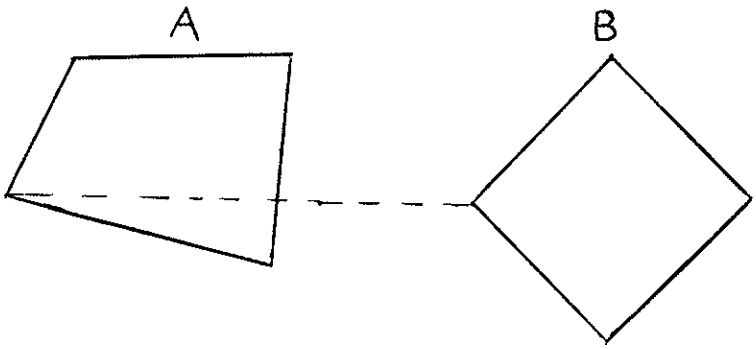
We are also interested in the distance between compact sets. Let $(\mathbf{X}, d)$ be a complete

metric space. Let $A, B \in \mathcal{H}(\mathbf{X})$. Define the distance between A and B to be:

$$d(A,B) = max\{d(x,B) : x \in A\}.$$

This can be written as:

$$d(A,B) = max\{min\{d(x,y) : y \in B\} : x \in A\}.$$



Note that there are always points $x_0 \in A$ and $y_0 \in B$ such that $d(A,B) = d(x_0, y_0)$. Working through a problem is a useful means to understand this distance function. Let $(\mathbf{X}, d)$ be a complete metric space. Show that if A, B, and $C \in \mathcal{H}(\mathbf{X})$ then

$$B \subset C \Rightarrow d(A,C) \leq d(A,B).$$

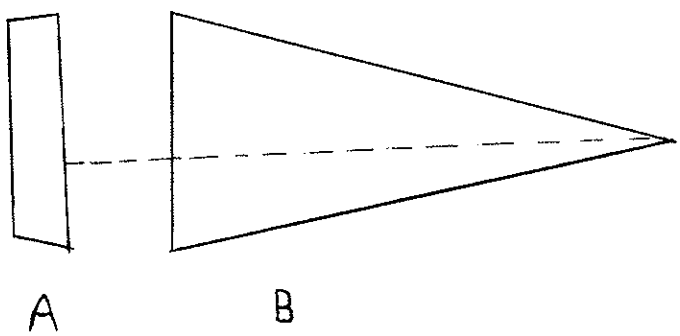
If $d(A,B) < d(A,C)$, then there exists a $y \in B$ such that

$$max\{min\{d(x,y) : y \in B\} : x \in A\} < max\{min\{d(x,z) : z \in C\} : x \in A\}.$$

But since $B \subset C$, there are no elements in B which are not in C, and so $d(A,B) \not< d(A,C)$. Hence, $d(A,C) \leq d(A,B)$.

Definition [Barnsley, p.34]: The *Hausdorff distance* between points A and B in $\mathcal{H}(\mathbf{X})$ is defined to be:

$$h_d(A,B) = max\{d(A,B), d(B,A)\}.$$



We will prove that h is a metric on the space $\mathcal{H}(\mathbf{X})$. Let $A, B, C \in \mathcal{H}(\mathbf{X})$.

1. $h_d(A,A) = max\{d(A,A), d(A,A)\} = d(A,A) = max\{d(x,A) : x \in A\} = 0.$
2. $0 < h_d(A,B) < \infty$ since $h_d(A,B) = d(a,b)$ for some $a \in A$ and $b \in B$, and d is a metric on $\mathbf{X}$.
3. $h_d(A,B) = h_d(B,A)$ is clear from the definition.
4. To show that $h_d(A,B) \leq h_d(A,C) + h_d(C,B)$, we first demonstrate that $d(A,B) \leq d(A,C) + d(C,B)$.

For any $a \in A$, we have:

$$d(a,B) = min\{d(a,b)|b \in B\} \leq min\{d(a,c) + d(c,b)|b \in B\}.$$

Since this is true for any c in C, we may write

$$d(a,B) \leq min\{d(a,c)|c \in C\} + max\{min\{d(c,b)|b \in B\} \ \forall c \in C.$$

Since this is true for any a in A, we have

$$d(a,B) \leq d(a,C) + d(C,B), \ \forall \ a \in A,$$

which can be written as

$$d(A,B) \leq d(A,C) + d(C,B).$$

Similarly, $d(B,A) \leq d(B,C) + d(C,A)$. Therefore

$$h_d(A,B) = max\{d(A,B), d(B,A)\}$$

$$\leq max\{d(C,B), d(C,B)\} + max\{d(A,C), d(C,A)\}$$

$$= h_d(B,C) + h_d(A,C). \quad \text{q.e.d.}$$

Our next goal will be to demonstrate that the space $\mathcal{H}(\mathbf{X})$ is complete. The need for completeness arises when we define what a fractal is. This guarantees that it will exist when we construct it using this method. Here is a definition and two lemmas which will aid in that proof:

Definition [Barnsley, p. 35]: Let $S \subset \mathbf{X}$, $\gamma > 0$. The *dilation* of S by a ball of radius γ is defined as

$$S + \gamma = \{y \in \mathbf{X} \mid d(x,y) \leq \gamma \text{ for some } x \in S\}.$$

Lemma 1 [Barnsley, p. 35]: Let A and B belong to $\mathcal{H}(\mathbf{X})$ where $(\mathbf{X}, d)$ is a metric space. Let $\epsilon > 0$. Then

$$h_d(A,B) \leq \epsilon \iff A \subset B + \epsilon \ \textit{and} \ \ B \subset A + \epsilon.$$

To prove this, we first show that $d(A,B) \leq \epsilon \Longleftrightarrow A \subset B + \epsilon$.
$\Longrightarrow$ Suppose $d(A,B) \leq \epsilon$. Then $\text{Max}\{d(a,B) : a \in A\} \leq \epsilon$ and $d(a,B) \leq \epsilon$ for all $a \in A$. So, for each $a \in A$, we have $a \in B + \epsilon$. Hence, $A \subset B + \epsilon$.
$\Longleftarrow$ Suppose $A \subset B + \epsilon$. Then for every $a \in A$, $d(a,b) \leq \epsilon$ for some $b \in B$. Hence, $d(a,B) \leq \epsilon$ for every a, and $d(A,B) \leq \epsilon$. So

$$d(A,B) \leq \epsilon \Longleftrightarrow A \subset B + \epsilon.$$

Similarly, it can be shown that $d(B,A) \leq \epsilon \Longleftrightarrow B \subset A + \epsilon$. Since $h_d(A,B) = \max$ $\{ d(B,A), d(A,B)\}$, $h_d(A,B) \leq \epsilon$ iff $d(A,B)$ and $d(B,A)$ are $\leq \epsilon$. These are in turn equivalent to $A \subset B + \epsilon$ and $B \subset A + \epsilon$. q.e.d.

Lemma 2 [Barnsley, p. 36]: Let (**X**,d) be a metric space. Let $\{A_n : n = 1, 2, \cdots, \}$ be a Cauchy sequence of points in $(\mathcal{H}(\mathbf{X}), h_d)$. Let $\{n_j\}_{j=1}^{\infty}$ be an infinite sequence of integers $0 < n_1 < n_2 < \cdots$. Suppose that we have a Cauchy sequence $\{x_{nj} \in A_{nj} : j = 1, 2, 3, \ldots\}$ in $(\mathbf{X}, d)$. Then there exists a Cauchy sequence $\{\tilde{x}_n \in A_n : n = 1, 2, 3, \ldots\}$ such that $\tilde{x}_{nj} = x_{nj}$ for all $j = 1, 2, 3, \ldots$.

What this lemma says is that, given a Cauchy sequence of compact sets (sets which get closer and closer to each other) and a Cauchy sequence of points in those sets (though not necessarily one point from each consecutive set), that there does in fact exist a Cauchy sequence which can be constructed and includes every set. To sketch the proof, we construct a sequence by choosing $\tilde{x}_n$ to be the closest point in A_n to x_{nj}, where $n \in \{n_{j-1}, ..., n_j\}$. It can be shown that this new sequence is indeed a Cauchy sequence from the fact that $\{A_n\}$ and $\{x_{nj}\}$ are both Cauchy.

The Completeness Theorem of $(\mathcal{H}(\mathbf{X}), h_d)$ [Barnsley, p. 37]: Let $(\mathbf{X}, d)$ be a complete metric space. Then $(\mathcal{H}(\mathbf{X}), h_d)$ is a complete metric space. Also, if $\{A_n \in \mathcal{H}(\mathbf{X})\}_{n=1}^{\infty}$ is a Cauchy sequence then

$$A = \lim_{n \to \infty} A_n \in \mathcal{H}(\mathbf{X})$$

exists and

$$A = \text{the limit points of all the convergent Cauchy sequences in } A_n.$$

This very important theorem forms the basis from which we can define what a fractal is, namely the attractor A. It is necessary for the space to be complete in order for the attractor to exist. The following is a sketch of the proof, which is broken into five parts:

1. $A \neq \emptyset$;
2. A is closed;
3. for any $\epsilon > 0$ there is a N such that for $n \geq N$, $A \subset A_n + \epsilon$;

4. A is totally bounded and (since closed) compact;

5. $Lim_{n\to\infty} A_n = A$.

First we show that A is non-empty. This is done by finding a Cauchy sequence $\{a_i \in A_i\}$ in $\mathbf{X}$. Since $\{A_n\}$ forms a Cauchy sequence, we find a sequence of positive integers $N_1 < N_2 < \cdots < N_n < \cdots$ such that

$$h_d(A_m, A_n) < \frac{1}{2^i} \quad for \;\; m, n > N_i.$$

Pick $x_{Ni} \in A_{Ni}$ for which $d(x_{Ni-1}, x_{Ni}) \leq \frac{1}{2^{i-1}}$. In this manner we can find an infinite sequence such that $d(x_{Ni}, x_{Ni+1}) \leq \frac{1}{2^i}$. This sequence $\{x_{Ni}\}$ can be shown to be a Cauchy sequence in $\mathbf{X}$. Let $\epsilon > 0$ be given. Pick $N(\epsilon)$ such that $\sum_{i=N(\epsilon)}^{\infty} \frac{1}{2^i} < \epsilon$. Then we have for $m > n \geq N(\epsilon)$

$$d(x_{Nm}, x_{Nn}) \leq d(x_{Nm}, x_{Nm+1}) + \cdots + d(x_{Nn-1}, x_{Nn})$$

$$< \sum_{i=N(\epsilon)}^{\infty} \frac{1}{2^i} < \epsilon.$$

By Lemma 2, we know that there is a convergent subsequence $\{a_i \in A_i\}$ for which $a_{Ni} = x_{Ni}$. Hence the limit of this sequence exists and is in A, making it nonempty.

We now move on to show that A is closed. Suppose we have a sequence of a_i's $\in A$ that converges to a point a. If $a \in A$, then A is closed. We have for each i a sequence $\{x_{in}\} \in A_n$ such that $Lim_{n\to\infty}\{x_{in}\} = a_i$. There exists an increasing sequence of positive integers $\{N_i\}$with $d(a_{Ni}, a) < \frac{1}{i}$. There is also a subsequence of integers $\{m_i\}$ such that $d(x_{Ni,mi}, a_{Ni}) \leq \frac{1}{i}$. Hence, $d(x_{Ni,mi}, a) \leq \frac{2}{i}$. If we let $y_{mi} = x_{Ni,mi}$, then $y_{mi} \in A_{mi}$ and $Lim_{i\to\infty} y_{mi} = a$. By Lemma 2, $\{y_{mi}\}$, a Cauchy sequence, can be extended to a convergent sequence $\{z_i \in A_i\}$. Hence, by definition of A, $a \in A$. Therefore, A is closed.

Now we show that, for large n, $A \subset A_n + \epsilon$, where $\epsilon > 0$ is given. So, we know that for any $\epsilon > 0$, there is a N such that for $m, n > N$, $h(A_m, A_n) \leq \epsilon$. If $m \geq n$, $A_m \subset A_n + \epsilon$. (This is from lemma 1). We then pick $a \in A$. We examine the sequence $a_i \in A_i$ which converges to a, and note that N can be large enough to make, for $m \geq N$, $d(a_m, a) < \epsilon$. By the compactness of A_n, $A_n + \epsilon$ is closed. Since $A_m \subset A_n + \epsilon$, $a_m \subset A_n + \epsilon$ for all $m \geq N$. Hence $a \in A_n + \epsilon$ and $A \subset A_n + \epsilon$, for large n.

We need to introduce a definition here, in order to do the following part of the proof.

Definition [Barnsley, p. 20]: Let $S \subset \mathbf{X}$ be a subset of a metric space $(\mathbf{X}, d)$. S is *totally bounded* if, for each $\epsilon > 0$, there is a finite set of points $\{y_1, y_2, \ldots, y_n\} \subset S$

such that whenever $x \in S$, $d(x, y_i) < \epsilon$ for some $y_i \in \{y_1, y_2, \ldots, y_n\}$. This set of points is called an $\epsilon - net$.

It will now be shown that A is totally bounded and since we have seen that it is closed, is compact. We take a contrapositive approach: suppose A were not totally bounded. Then for some $\epsilon > 0$ there is no finite ϵ-net. We could then find a sequence $\{x_i\}_{i=1}^{\infty}$ in A such that $d(x_i, x_j) \geq \epsilon$ for $i \neq j$. We know that for large n, $A \subset A_n + \frac{\epsilon}{3}$. For each x_i, there is a corresponding $y_i \in A_n$ such that $d(x_i, y_i) \leq \frac{\epsilon}{3}$. We know that each A_i is compact, however, so there is a convergent subsequence from which we can find points as close together as we wish. We pick two such points y_{ni} and y_{nj} such that the distance between them is $\frac{\epsilon}{3}$. But by the quadrangle inequality,

$$d(x_{ni}, x_{nj}) \leq d(x_{ni}, y_{ni}) + d(y_{ni}, y_{nj}) + d(y_{nj}, x_{nj}) < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.$$

Therefore, our conjecture that there is no finite ϵ-net must be false, and A is indeed totally bounded. By the fact that it is closed and totally bounded, it is compact.

Finally, we need to show that A_n actually does approach A in the limit. By lemma 1 and the fact that $A \subset A_n + \epsilon$, we must now show that $A_n \subset A + \epsilon$ for A_n and A to be moving within an ϵ distance from each other as n grows. Let $\epsilon > 0$ and let N satisfy for $m, n > N$, $h_d(A_m, A_n) \leq \frac{\epsilon}{2}$. Then for $m, n \geq N, A_m \subset A_n + \frac{\epsilon}{2}$. Let $y \in A_n$. There is an increasing sequence of positive integers: $n < N_1 < N_2 < N_3 < \cdots < N_k < \cdots$ which for $m, k \geq N_j$ yield $A_m \subset A_k + \frac{\epsilon}{2^{j+1}}$. Now since $A_n \subset A_{N1} + \frac{\epsilon}{2}$, with $y \in A_n$, there is an $x_{N1} \in A_{N1}$ such that $d(y, x_{N1}) \leq \frac{\epsilon}{2}$. Continuing, we find a sequence of x_{Ni}'s such that $d(x_{Nj}, x_{Nj+1}) < \frac{\epsilon}{2^{j+1}}$. Through the $j - angle$ inequality, we see that

$$d(y, x_{Nj}) \leq d(y, x_{N1}) + \cdots + d(x_{Nj-1}, x_{Nj}) \leq \frac{\epsilon}{2} + \cdots + \frac{\epsilon}{2^j} < \epsilon,$$

for all j. So, $\{x_{Nj}\}$is a Cauchy sequence which converges to a point $x \in A$. Also, $d(y, x_{Nj}) \leq \epsilon$ implies that $d(y, x) \leq \epsilon$. Hence, $A_n \subset A + \epsilon$ for $n \geq N$.
We have shown that Lim $A_n = A$ and also that $(\mathcal{H}(\mathbf{X}), h_d)$ is a complete metric space. q.e.d.

We now proceed to examine transformations on metric spaces. They are the building blocks of fractals. First let's define what exactly a transformation is.

Definition [Barnsley, p. 43]: Let $(\mathbf{X}, d)$ be a metric space. A *transformation* on $\mathbf{X}$ is a function $f : \mathbf{X} \to \mathbf{X}$.

Definition [Barnsley, p. 44]: Let $f : \mathbf{X} \to \mathbf{X}$ be a transformation on a metric space. The *forward iterates* of f are defined as follows. $f^{o1}(x) = f(x)$, and if $f^{on}(x)$, the nth iterate of f, is defined, then $f^{o(n+1)}(x) = f(f^{on}(x))$ for $n = 1, 2, 3, \ldots$. If f is invertible then the *backward iterates* of f are similarly defined. $f^{o(-1)}(x) = f^{(-1)}(x)$, and if $f^{o(-m)}(x)$, the mth iterate of f^{-1}, is defined, then $f^{o(-m-1)}(x) = f^{-1}(f^{o(-m)}(x))$, for

$m = 1, 2, 3, \cdots$.

An example of a simple transformation is $f : [0,1] \rightarrow [0,1]$ defined by $f(x) = 4x(1-x)$ This transformation is not *one-to-one*, since $f(0) = f(1)$, however it is *onto*. It is not invertible since it is not one-to-one. We will deal primarily with affine transformations on $\boldsymbol{R}^2$. They are defined as follows:

Definition [Barnsley, p. 50]: A transformation $w : \boldsymbol{R}^2 \rightarrow \boldsymbol{R}^2$ of the form

$$w(x_1, x_2) = (ax_1 + bx_2 + e, cx_1 + dx_2 + f)$$

where a, b, c, d, e, and f are real numbers, is called a 2-D *affine* transformation. Affine transformations have the equivalent notations

$$w(x) = w\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix} = Ax + t.$$

The affine transformation $Ax + t$ thus first deforms space relative to the origin, and secondly translates space according to the vector t. Let's look at an example. Suppose we want to find the affine transformation which takes the triangle with vertices at (1,0), (0,1), and (0,0) to the triangle with vertices at (-1,2),(4,5), and (3,0).
Our matrix A looks like

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} -4 & 1 \\ 2 & 5 \end{pmatrix}$$

and

$$t = \begin{pmatrix} e \\ f \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \end{pmatrix}.$$

The transformation matrix

$$\begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix}$$

represents a counter-clockwise rotation. The matrix

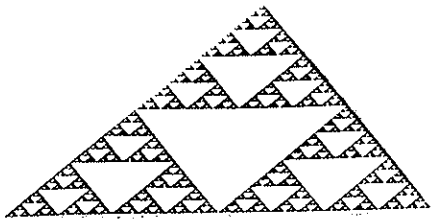
$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

is the form of a reflection over the x-axis, and

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

serves as a reflection about the y-axis.

A well known fractal is the Sierpinski Triangle, which is a triangle with infinite copies of itself inside it:



We will denote a Sierpinski triangle by $\mathcal{S}$ One example of an affine transformation from $\mathcal{S}$ with vertices at (0,0), (1,0), and $(\frac{1}{2}, 1)$ such that

$$w(\mathcal{S}) \subset \mathcal{S}, \quad with \quad w(\mathcal{S}) \neq \mathcal{S}$$

where $x = (x_1, x_2)$, is

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

This will take the entire triangle and move it into the next largest one closest to the origin.

An interesting transformation which is of special significance in the study of chaotic dynamical systems is called the *shift transformation* and is defined from

$$T : \Sigma \to \Sigma \text{ such that } T(x_1x_2x_3\cdots) = x_2x_3x_4\cdots$$

Of critical importance in defining fractals is the notion of points which are invariant under a transformation. On a fixed point, the transformation acts as follows:

$$f(x_f) = x_f$$

Fixed points limit the effect of the transformation by creating a pivot point to which the space under the transformation is affixed. An example of a fixed point is useful in understanding their importance. Given the transformation

$$f(x) = ax + b \quad a \neq 0, \quad a \neq 1, \quad a, b \in \boldsymbol{R},$$

we find the fixed point, x_f, by inserting x_f into the equation for x and solving. Clearly, $x_f = \frac{b}{1-a}$. The fixed points of a shift transformation are obviously $x_f = 1111111...$, $x_f = 2222222...$, etc., since shifting it over will produce the same point. q.e.d.

As will be seen, it is of paramount importance to discuss the idea of a transformation which, when applied, shrinks the space toward some point. This notion is formalized

in the definition of a contraction mapping.

Definition [Barnsley, p. 75]: A transformation $f : \mathbf{X} \to \mathbf{X}$ on a metric space $(\mathbf{X}, d)$ is called a *contraction mapping* if there is a constant $0 \leq s < 1$ such that

$$d(f(x), f(y)) \leq \ s \cdot d(x, y) \ \ \forall \ x, y \in \mathbf{X}.$$

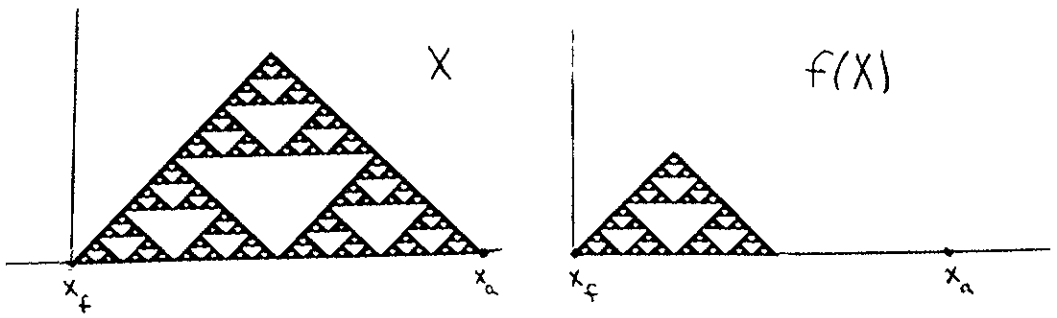
We call s the contractivity factor. A consequence of this definition may help illustrate the contraction map idea. We will soon show that every contraction mapping has a unique fixed point. For the moment, let's assume it is true. Let $(\mathbf{X}, d)$ be a compact space which contains more than one point. Let $f : \mathbf{X} \to \mathbf{X}$ be a contraction mapping. Show $f(\mathbf{X}) \subset \mathbf{X}$ but $f(\mathbf{X}) \neq \mathbf{X}$. Since f maps $\mathbf{X}$ to itself, then obviously $f(\mathbf{X}) \subset \mathbf{X}$. Since $\mathbf{X}$ is compact, it is totally bounded. So take

$$max\{d(x_f, x) \ \mid \ x \in \mathbf{X}\} = a_1.$$

Let the x which satisfies the above equation be called x_a. By the defintion of contractivity,

$$max\{d(f(x_f), f(x)) \ x \in \mathbf{X}\} \ \ < \ \ a_1.$$

So if x_a were in $f(\mathbf{X})$, then it would be further than the greatest distance of any point from the fixed point. This is obviously a contradiction. Hence, x_a is not in $f(\mathbf{X})$.



The Contraction Mapping Theorem [Barnsley, p. 76]: Let $f : \mathbf{X} \to \mathbf{X}$ be a contraction mapping on a complete metric space $(\mathbf{X}, d)$. Then f posesses exactly one fixed point, $x_f \in \mathbf{X}$, and moreover, for any $x \in \mathbf{X}$, the sequence $\{f^{on}(x) : n = 0, 1, 2, 3, \cdots\}$ converges to x_f. Hence,

$$\lim_{n \to \infty} f^{on}(x) \ \ = \ \ x_f \ \ \forall x \in \mathbf{X}.$$

For the proof, let $x \in \mathbf{X}$, and let s be the contractivity factor of f. Then, for fixed x,

$$d(f^{on}(x), f^{om}(x)) \leq s^{min\{m,n\}} \cdot d(x, f^{o|n-m|}(x))$$

for all m, n = 0,1,2,.... In particular, for k = 0,1,2,...,

$$d(x, f^{ok}(x)) \leq d(x, f(x)) + d(f(x), f^{o2}(x)) + \cdots + d(f^{o(k-1)}(x), f^{ok}(x))$$

$$\leq (1 + s + s^2 + \cdots + s^{k-1}) \cdot d(x, f(x))$$

$$\leq \left(\frac{1}{1-s}\right) \cdot d(x, f(x)).$$

So we now can say that

$$d(f^{om}(x), f^{on}(x)) \leq \frac{s^{min\{n,m\}}}{1-s} d(x, f(x)).$$

So, as n and m grow, $d(f^{om}(x), f^{on}(x))$ gets arbitrarily small and thus $\{f^{on}(x)\}_{n=0}^{\infty}$ is a Cauchy sequence. By $\mathbf{X}$'s completeness, there is an $x_f \in \mathbf{X}$ to which our sequence converges.
We now show that x_f is a fixed point. Since f is contractive it is continuous. To see this, for $\epsilon > 0$, pick $\delta = \frac{\epsilon}{s}$. Then

$$d(x, y) < \delta$$

implies

$$d(f(x), f(y)) < s \cdot d(x, y) < s \cdot \frac{\epsilon}{s} = \epsilon.$$

Therefore it is perfectly legitimate to say that

$$f(x_f) = f(\lim_{n\to\infty} f^{on}(x)) = \lim_{n\to\infty} f^{o(n+1)}(x) = x_f.$$

There is only one such point, x_f. Suppose there existed another, say y_f. Then

$$d(x_f, y_f) = d(f(x_f), f(y_f)) \leq s \cdot d(x_f, y_f).$$

But this implies that $(1-s) \cdot d(x_f, y_f) \leq 0$, and hence $d(x_f, y_f) = 0$ and thus $x_f = y_f$. q.e.d.

Here is an example of a transformation which turns out to be a contraction mapping. Let $f : \boldsymbol{R} \to \boldsymbol{R}$ be $f(x) = \frac{1}{2}x + \frac{1}{2}$. To verify that f is a contraction mapping, we find an s such that:

$$d(f(x), f(y)) \leq s \cdot d(x, y) \ \forall x, y \in \boldsymbol{R}.$$

Let $s = \frac{1}{2}$. Then if $d(x, y) = \gamma$,

$$d(\frac{1}{2}x + \frac{1}{2}, \frac{1}{2}y + \frac{1}{2}) = d(\frac{1}{2}x, \frac{1}{2}y) = \frac{1}{2}d(x, y)$$

What point do all the points converge to? To answer this, we find

$$\lim_{n\to\infty} f^{on}(x).$$

Take $x = 0$. Then $f(x) = \frac{1}{2}$, $f^2(x) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$, $f^3(x) = \frac{3}{8} + \frac{1}{2} = \frac{7}{8}$. Continuing on in this fashion yields the sequence

$$\frac{1}{2}, \frac{3}{4}, \frac{7}{8}, \frac{15}{16}, \frac{31}{32}, \cdots$$

It is easily seen that as this continues, the limit is

$$\frac{1}{2}(1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots).$$

This is none other than the geometric series converging to 1. It can be verified that 1 is the fixed point, since $\frac{1}{2}(1) + \frac{1}{2} = 1$. q.e.d.

An example of a contraction mapping defined over an equilateral $\mathcal{S}$ with vertice at the origin and at (1,0) is as follows:

$$w(x) = \begin{pmatrix} \frac{1}{2}cos120^o & -\frac{1}{2}sin120^o \\ \frac{1}{2}sin120^o & \frac{1}{2}cos120^o \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}.$$

The fixed point in this mapping solves the equation

$$\begin{pmatrix} -\frac{1}{4} & -\frac{\sqrt{3}}{4} \\ \frac{\sqrt{3}}{4} & -\frac{1}{4} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

When solved,

$$x_f = \begin{pmatrix} \frac{5}{14} \\ \\ \frac{\sqrt{3}}{14} \end{pmatrix}.$$

This means that repeated applications of the transformation will eventually take the entire set $\mathcal{S}$ towards one point.

Now we have the tools to define what fractals really are. A *deterministic fractal* is a fixed point of a contractive transformation on $(\mathcal{H}(\mathbf{X}), h_d)$. We shall develop this idea as we progress. First we need to observe some important lemmas.

Lemma 3 [Barnsley, p. 80]: Let $w : \mathbf{X} \to \mathbf{X}$ be a continuous mapping on any metric space $(\mathbf{X}, d)$. Then w maps $\mathcal{H}(\mathbf{X})$ into itself.

The proof of this involves the preservation of convergent subsequences of the known infinite sequences in some compact subset of $\mathbf{X}$ by the continuity of our transformation.

Lemma 4 [Barnsley, p. 80]: Let $w : \mathbf{X} \to \mathbf{X}$ be a contraction mapping on the metric space $(\mathbf{X}, d)$ with contractivity factor s. Then $w : \mathcal{H}(\mathbf{X}) \to \mathcal{H}(\mathbf{X})$ defined by

$$w(B) = \{w(x) | x \in B\} \text{ for all } B \in \mathcal{H}(\mathbf{X})$$

is a contraction mapping on $(\mathcal{H}(\mathbf{X}), h_d)$ with contractivity factor s.

We have shown that since w is a contraction mapping, it is continuous, so by Lemma 3, w maps $\mathcal{H}(\mathbf{X})$ into itself. So, pick $B, C \in \mathcal{H}(\mathbf{X})$. Then

$$d(w(B), w(C)) = max\{min\{d(w(x), w(y))|y \in C\}|x \in B\}$$

$$\leq max\{min\{s \cdot d(x, y)|y \in C\}|x \in B\} = s \cdot d(B, C).$$

Similarly, $d(w(C), w(B)) \leq s \cdot d(C, B)$. Therefore,

$$h_d(w(B), w(C)) = max\{d(w(B), w(C)), d(w(C), w(B))\} \leq s \cdot h_d(B, C).$$ q.e.d.

This tells us how contraction mappings act on $\mathcal{H}(\mathbf{X})$. We now make a critical step in our understanding of contraction mappings. We may now form them as the union of several contractive transformations.

Lemma 5 [Barnsley, p. 81]. Let $(\mathbf{X}, d)$ be a metric space. Let $\{w_n \mid n = 1, 2, ..., N\}$ be contraction mappings on $(\mathcal{H}(\mathbf{X}), h_d)$ with contractivity factor s_i for each w_i. Define $W : \mathcal{H}(\mathbf{X}) \to \mathcal{H}(\mathbf{X})$ by

$$W(B) = w_1(B) \cup w_2(B) \cup \cdots \cup w_n(B) = \bigcup_{n=1}^{N} w_n(B)$$

for each $B \in \mathcal{H}(\mathbf{X})$. Then W is a contraction mapping with contractivity factor $s = max\{s_n \mid n \in 1, 2, ..., N\}$.

An essential part of our study is the notion of an iterated function system (IFS).

Definition [Barnsley, p. 82]: An *iterated function system* consists of a complete metric space together with a finite set of contraction mappings. The notation for such an IFS is $\{\mathbf{X}; w_n, n = 1, 2, ..., N\}$.

This is sometimes referred to as a "hyperbolic IFS" , since it contracts, but we often drop the "hyperbolic". The major theorem which explains how to find deterministic fractals is the following, which develops the theme of the fixed point of a contraction mapping and introduces the idea of an *attractor.*

Theorem [Barnsley, p. 82]: Let $\{\mathbf{X}; w_n, n = 1, 2, ..., N\}$ be an IFS with contractivity factor s. Then the transformation $W : \mathcal{H}(\mathbf{X}) \to \mathcal{H}(\mathbf{X})$ defined by

$$W(B) = \bigcup_{n=1}^{N} w_n(B)$$

for all $B \in \mathcal{H}(\mathbf{X})$, is a contraction mapping on the complete metric space $(\mathcal{H}(\mathbf{X}), h_d)$ with contractivity factor s. So

$$h(W(B), W(C)) \leq s \cdot h_d(B, C)$$

for all $B, C \in \mathcal{H}(\mathbf{X})$. Its unique fixed point, $A \in \mathcal{H}(\mathbf{X})$, obeys

$$A = W(A) = \bigcup_{n=1}^{N} w_n(A),$$

and is given by $A = \lim_{n\to\infty} W^{on}(B)$ for any $B \in \mathcal{H}(\mathbf{X})$.

Definition [Barnsley, p. 82]: The fixed point A described above is called the *attractor* of the IFS.

The following is another example of a contraction mapping, this time involving code space. The classical Cantor Set, which consists of the closed unit interval with deleted open middle thirds (continued ad infinitum), can be thus defined as the limit of a nested sequence of closed sets and can be constructed using two contraction mappings. Let

$$w_1 = \frac{1}{3}x, \quad w_2 = \frac{1}{3}x + \frac{2}{3}$$

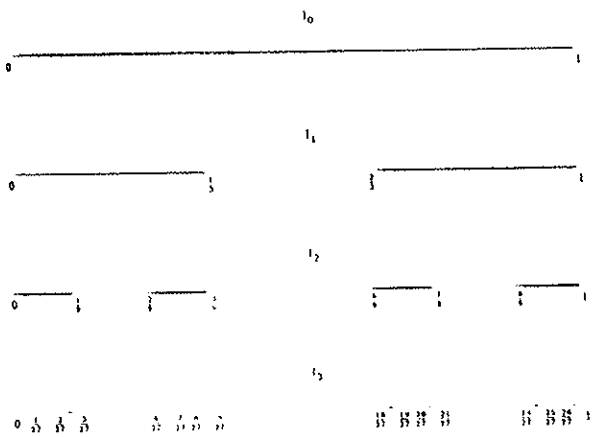
The IFS $\{[0,1]; w_1, w_2\}$ creates the desired set. To see this, observe that

$$w[0,1] = w_1[0,1] \cup w_2[0,1] = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1].$$

This is the first tier of the construction. When the mappings are applied again, we get

$$w^{o2}[0,1] = w_1[0, \frac{1}{3}] \cup w_2[0, \frac{1}{3}] \cup w_1[\frac{2}{3}, 1] \cup w_2[\frac{2}{3}, 1]$$

It is easy to see that if we continue this process ad infinitum, we have the classical ternary Cantor Set, which is a fractal. It is also neatly defined in the ternary system as all numbers consisting of 0's and 2's. Note that we do have points with a 1 in them, such as $.1 = \frac{1}{3}$, but we could just have easily represented .1 as .02222222...



Let (Σ_3, d_c) be code space on three symbols. Then $d_c(x,y) = \sum_{n=1}^{\infty} \frac{|x_n - y_n|}{4^n}$. Define

$w_1, w_2 : \Sigma_3 \to \Sigma_3$ by

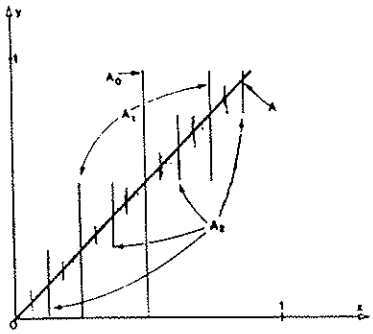
$$w_1(x) = 0x_1x_2x_3\cdots$$

and

$$w_2(x) = 1x_1x_2x_3\cdots$$

It can be shown that they are both contraction mappings on (Σ_3, d_c).
$w_1 : d(w_1(x), w_1(y)) = \sum_{n=0}^{\infty} \frac{|x_n - y_n|}{4^{n+1}} = \frac{1}{4} \cdot d(x,y)$. Similarly, $d(w_2(x), w_2(y)) = \frac{1}{4} \cdot d(x,y)$. The attractor of the IFS $\{\Sigma_3; w_1, w_2\}$ consists of all words in code space with 2's and 0's. It can be shown that this is metrically equivalent to a cantor subset of [0,1].

Another example of an IFS is: $\{\boldsymbol{R}^2; \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}\}$ Suppose we let $A_0 = \{(\frac{1}{2}, y) |\ 0 \le y \le 1\}$, and let $W^{on}(A_0) = A_n$. The attractor A is actually the line segment from the origin to the point (1,1).



Now we are ready to rethink our Sierpinski Triangle, and formulate it as the attractor of an IFS. If we have the IFS:

$$\{0 \le x \le 1, 0 \le y \le 1; w_1, w_2, w_3\}$$

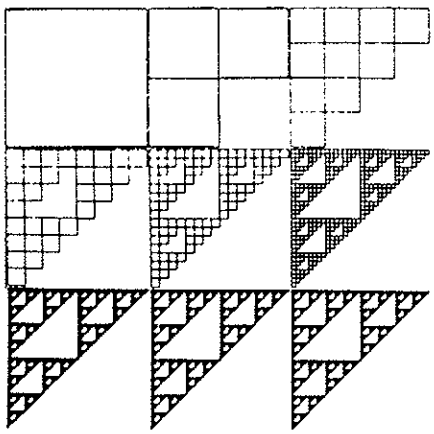
where

$$w_1 = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}, \quad w_2 = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix},$$

$$w_3 = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} \frac{1}{4} \\ \frac{1}{2} \end{pmatrix}.$$

The first mapping takes the unit square, and contracts it into the unit quartersquare. The second does the same but then translates it over by a half on the y-axis. The third shrinks it down and translates it up and over on top of the other two. It is fascinating to observe that whatever set you start with in the unit square, the iteration of these

contraction mappings always yields the exact same attractor. This holds whether the set is the whole square, an flat elephant, or only one point!



In light of how we defined our contraction mapping as the union of a finite number of contraction mappings acting on compact subsets, I was naturally curious as to what happened in the infinite case, i.e. do they also form contraction mappings? I discovered a counterexample which says that for at least one case, this does not hold. Consider the IFS: $\{\mathbf{X} : w_1, w_2, ...\}$ where

$$w_n(x) = \frac{2^n - 1}{2^n}x + \frac{2^n - 1}{2^n}.$$

Thus, given $B \in \mathrm{H}(\mathbf{X})$,

$$W(B) = \bigcup_{n=1}^{\infty} w_n(B).$$

Then if we examine the mapping applied to two intervals, [1,2] and [2,3], we see that:

$$W[1,2] = [\frac{1}{2}\cdot 2, \frac{1}{2}\cdot 3] \cup [\frac{3}{4}\cdot 2, \frac{3}{4}\cdot 3] \cup [\frac{7}{8}\cdot 2, \frac{7}{8}\cdot 3] \cup \cdots = [1,3).$$

$$W[2,3] = [\frac{1}{2}\cdot 3, \frac{1}{2}\cdot 4] \cup [\frac{3}{4}\cdot 3, \frac{3}{4}\cdot 4] \cup [\frac{7}{8}\cdot 3, \frac{7}{8}\cdot 4] \cup \cdots = [\frac{3}{2}, 4).$$

We replace our max by sup (since these sets are not compact) to denote the least upper bound, and observe that

$$d(W[1,2], W[2,3]) = sup\{d(x, [\frac{3}{2}, 4))|x \in [1,3)\} = \frac{1}{2}$$

and

$$d(W[2,3], W[1,2]) = sup\{d(x, [1,3))|x \in [\frac{3}{2}, 4)\} = 1.$$

Hence, $h_d(W[1,2], W[2,3]) = 1$. But since $h_d([1,2],[2,3]) = 1$, we see that the mapping is not a contraction mapping.

I hope the reader feels the same joy and wonder that I do when learning how to make fractals using IFS's. This field is young, and awaits full development, but I think that the near future will witness a renaissance in information storage, transmittal, and retrieval. The ability to communicate such complicated shapes with so little information is a technology waiting to be used.

Fractals and their accompanying mathematics are going through some tough times. There are mathematicians who despise anything connected with this field, and merely discard them as pretty pictures with no substance. My only response to this is to ask those people to think back to when people probably thought the same thing about the square, the triangle, and the icosahedron. The ancients probably felt that there was some great secrets about their shape that they couldn't understand. The incredible developments in abstract algebra and algebraic geometry have proven that there are immense and wonderful structures underlying these classical shapes. Fractal mathematicians feel the same about these new objects, which are marvelous to behold and require only a bit of imagination to unlock their secrets.

Bibliography

Barnsley, Michael F., *Fractals Everywhere*, Harcourt Brace Jovanovich, San Diego, 1988.

Kirkwood, James R., *An Introduction to Analysis*, PWS-Kent, Boston, 1989.