

A NOTE TO THE READER

This paper is submitted as a ‘fagprojekt’ in conjunction with the seminar course entitled ‘Seminar in Information Theory and its Applications’ hosted by Flemming Topsøe at the Institute for Mathematical Sciences at University of Copenhagen in the fall of 2000. The paper is not related to any specific presentation made during the seminar, but rather is the result of independent reading of [Shannon, 1948], [Immink et al., 1998] and [Falconer, 1990] and informal discussions with a number of the participants of the seminar.

The primary goal of the paper is to follow a lead about the connection between crosswords and entropy that was put forward by Shannon in 1948. The secondary goal of the paper is to try out three different styles of mathematical writing¹. Thus the first part of the paper is an attempt to deal with the calculation of the number of crosswords in a way that require minimal mathematical theory. The second part of the paper is a exposition of mathematical content in the traditional way. The last part is intended to be a report of what might be termed a ‘numerical experiment’ and its background.

Peter Andreasen
Copenhagen, November 2001

¹On the importance of Mathematical Writing, Paul Halmos says: “Mathematicians who merely *think* great theorems have no more done their job than painters who merely *think* great painting.” So it seems reasonable to train not only the skills of presenting mathematical content, but also the ability to present in different forms.

Crosswords and Hausdorff dimensions of self-similar sets

By PETER ANDREASEN

November 2001

Abstract

This paper is organized in three almost self contained sections. The first section is a popular discussion on the number of crosswords based on a remark of Shannon made in [Shannon, 1948]. Although never published, the 'back-of-the-envelope' calculation we present is believed to be the same argument Shannon had in mind when he wrote the passage on crosswords.

The second section presents the Hausdorff dimension and the box-counting dimension and gives an introduction to the theory of self similar sets. Methods for calculating the Hausdorff dimension of self similar sets are discussed; in particular we prove Hutchinson's formula for the dimension of sets represented by iterated function systems (IFS).

The third and last section describes how the topics of the two previous sections are in fact connected. Indeed, we argue that the problem of calculating the number of crosswords is related to finding the Hausdorff dimension of certain self similar sets. In addition we identify a connection between box counting dimension and Hartley entropy.

I ON THE NUMBER OF CROSSWORDS

The American mathematician Claude E. Shannon is widely recognized as the father of information theory. His most famous paper is "A Mathematical Theory of Communication" from 1948 in which he lays the foundation for the modern information theory, and even today the paper shows an impressive combination of clarity and vision. We quote a few passages from the end of section 7, wherein we find a curious reference to such mundane matters as crossword puzzles...

The ratio of the entropy of a source to the maximum value it could have while still restricted to the same symbols will be called its *relative entropy*. This, as will appear later, is the maximum compression possible when we encode into the same alphabet. One minus the relative entropy is the *redundancy*. The redundancy of ordinary English, not considering statistical structure over greater distances than about eight letters, is roughly 50%. This means that when we write English half of what we write is determined by the structure of the language and half is chosen freely. The figure 50% was found by several independent methods which all gave results in this neighborhood. One is by calculation of the entropy of the approximations to English. A second method is to delete a certain fraction of the letters and then let someone attempt to restore them. If they can be restored when 50% are deleted the redundancy must be greater than 50%. A third method depends on certain known results in cryptography.

Two extremes of redundancy in English prose are represented by Basic English and by James Joyce's book *Finnegans Wake*. The Basic English vocabulary is limited to 850 words and the redundancy is very high. This is reflected in the expansion that occurs when a passage is translated into Basic English. Joyce on the other hand enlarges the vocabulary and is alleged to achieve a compression of semantic content.

The redundancy of a language is related to the existence of crossword puzzles. If the redundancy is zero any sequence of letters is a reasonable text in the language and any two-dimensional array of letters forms a crossword puzzle. If the redundancy is too high the language imposes too many constraints for large crossword puzzles to be possible. A more detailed analysis shows that if we assume the constraints imposed by the language are of a rather chaotic and random nature, large crossword puzzles are just possible when the redundancy is 50%. If the redundancy is 33%, three-dimensional crossword puzzles should be possible, etc.

[Shannon, 1948]

If Shannon was right about the connection between the redundancy of a language and the possibility of crossword puzzles, then we should be able to gain insight into the notion of redundancy (and hence also entropy as it is just one minus the redundancy) by calculating the number of possible crosswords. At least in theory. We intend to try it out in practice.

What is a crossword, really?

Most people have solved a crossword puzzle or played Scrabble. The existence of word-games like those are not to be taken for granted, though. As we are going to see, the existence of crosswords is entirely at the mercy of the underlying language.

Before we proceed, we need to get a few definitions straight. First, what *is* a crossword? Let us take a look at one:

g	e	m	□
a	r	e	□
m	a	t	h
e	□	□	e

What we see are rows and columns of words (single letters are accepted as words) separated by white squares.¹ Now, the words in a crossword need not be English as in the example above. We might want to create a Danish crossword or we might even want to have the columns and rows be quotes from Shakespeare's sonnets. To be able to handle such complex rules for the creation of crosswords we make the following definition.

Definition 1 A language L is a set of sequences of letters from an alphabet A (say, the letters 'a' through 'z' and the symbol '□'). A crossword of size n is a matrix with the dimensions $n \times n$ where all of the rows are sequences (of length n) from L and all of the columns are sequences (of length n) from L .

If we want to make a really sophisticated crossword, we may let L be all the possible quotes from Shakespeare. In that case we should use an alphabet which included the letters as well as space and the various punctuation symbols. If we wanted to make an ordinary crossword, we would have L be equal to any sequence you can make by taking words from a dictionary and gluing them together with one or more □'s in between. Typically we would also accept single letters as valid words so that crosswords like the one shown above become legal (otherwise the bottom row "e□□e" would not be allowed). At any rate, the alphabet A would simply be the letters 'a'... 'z' and the special symbol □.

¹It is in the white squares you will normally find the hints needed to solve the puzzle – and in most crosswords the topmost row and the leftmost column are filled with these hints. For simplicity we will make no assumptions about the placement of the white squares.

How many are there?

It is obvious, that very small crosswords are easily constructed. It is also easy to create a some large (but quite dull) crosswords: If we keep alternating the rows between 'I□I□I...' and '□I□I□...', we certainly get a valid (and as large as we want) ordinary English crossword. We therefore want to consider not only the existence of big crosswords, but also check if there are many different of them. We are going to calculate the number of big crosswords now.

Assume we have chosen an alphabet A and a language L over which the crosswords must be made. We use the notation $|A|$ for the number of letters and symbols in the alphabet. Let us introduce the following number as well,

$$c_L(n) = \text{number of sequences from } L \text{ of length } n.$$

Thus, when constructing a square crossword of size $n \times n$ over the language L , there are $c_L(n)$ possible choices for the first row. We will now use a small trick and for a moment employ a bit of probability theory: If we picked a random sequence of n letters from A , what are the chance that we got a 'valid' row, that is, a sequence from L ? The answer is

$$\frac{c_L(n)}{|A|^n}$$

because there are $c_L(n)$ valid sequences and $|A|^n$ possible sequences. An example: suppose we wanted to create a normal English crossword. In one dictionary, there are 49 words of length 2, and as mentioned above, we will consider all 27 letters (recall, that we have extended the alphabet with the special symbol \square) as words. Valid sequences of length 2 are then ' $\square\square$ ', ' $\square a$ '..., ' $\square z$ ', ' $a\square$ '... ' $z\square$ ', and then the 49 regular words of length 2, 'ad', 'ah', 'al', and so forth. This yields a total of 102 sequences, that is, $c_L(2) = 102$. The number of *all* possible sequences of length 2 is $|A|^2 = 27 \times 27 = 729$. And so the probability of getting a valid sequence of length 2 would be $102/729 \sim 0.14$, in this example.

However, that was the probability of just *one* valid row. What about the rest? The probability of all n rows being valid equals the above probability multiplied with itself n times

$$\left(\frac{c_L(n)}{|A|^n}\right)^n = \frac{c_L(n)^n}{|A|^{n^2}}.$$

Now for the columns the situation is identical. And because the columns are as high as the rows are wide, the result is the same: The probability of all n columns being valid (that is, from L) also equals

$$\frac{c_L(n)^n}{|A|^{n^2}}.$$

Now we may calculate the probability of a randomly selected matrix of $n \times n$ letters from A being in fact a crossword: We want *both* its rows *and* its columns to be valid, so we multiply:

$$\left(\frac{c_L(n)^n}{|A|^{n^2}}\right)^2 = \frac{c_L(n)^{2n}}{|A|^{2(n^2)}}$$

We now return to our original question: How many (big) crosswords are there? Well, we know the probability of a randomly selected matrix of $n \times n$ letters being a crossword, and there are a total of $|A|^{n \times n} = |A|^{n^2}$ possible $n \times n$ matrices, so we may write

$$N_n = |A|^{n^2} \times \frac{c_L(n)^{2n}}{|A|^{2(n^2)}} = \frac{c_L(n)^{2n}}{|A|^{n^2}}, \quad (1)$$

defining N_n as the number of crosswords of size $n \times n$.

Explosive numbers

To get to the core of the matter, we need to do a few more calculations. First, we apply the logarithm to N_n :

$$\begin{aligned}\log N_n &= 2n \log c_L(n) - n^2 \log |A| \\ &= 2n^2 \left(\frac{\log c_L(n)}{n} - \frac{\log |A|}{2} \right) \\ &= \frac{2n^2}{\log |A|} \left(\frac{\log_{|A|} c_L(n)}{n} - \frac{1}{2} \right)\end{aligned}$$

The symbol $\log_{|A|}$ is used for the logarithm to the base of $|A|$, that is, $|A|^{\log_{|A|} x} = x$. Recall that we are interested in the number N_n when n grows large. In the expression above, the first fraction,

$$\frac{2n^2}{\log |A|},$$

just grows towards infinity as n does the same. The second fraction,

$$\alpha_n = \frac{\log_{|A|} c_L(n)}{n}$$

is more interesting (so we name it α_n). The value of $c_L(n)$ must be between 0 and $|A|^n$ (that should be clear from the definition of $c_L(n)$). So (assuming $c_L(n) > 0$) we see that $\log_{|A|} c_L(n)$ is between 1 and n . Thus, when n grows large, the value of α_n stays between 0 and 1. Let us assume that α_n in fact *converges*² to some number α between 0 and 1. We may now conclude, that if $\alpha < \frac{1}{2}$ the value of $\log N_n$ goes towards negative infinity ($-\infty$) as n becomes big. To see this very clearly, consider the formula from above (this time written using the symbol α_n , but otherwise identical):

$$\log N_n = \frac{2n^2}{\log |A|} \left(\alpha_n - \frac{1}{2} \right).$$

Assuming $\alpha < \frac{1}{2}$ then α_n will also be less than $\frac{1}{2}$ (provided n is big enough) and hence we find that the value of $(\alpha_n - 1/2)$ becomes negative, while the first fraction as mentioned above grows towards infinity.

If, on the other hand, $\alpha > \frac{1}{2}$, the value of $\log N_n$ approaches positive infinity (∞), provided we make n big enough.

Now consider the implications of α for N_n , the number of crosswords of size n . If $\log N_n$ is growing unlimited when n grows, then it indicates that N_n grows unlimited as well. If $\log N_n$ decreases to very large negative numbers when n grows, then N_n must be close to zero. Somehow the number α determines if the number of valid crosswords (N_n) gets very large as n (the size of the crosswords) gets large, *or* if the number of valid crosswords becomes almost zero!

It is rather clear, that the value α depends on L and only on L . Thus we will write α_L and say that the language L has α -value α_L .

²This is not a trivial assumption, but for languages of the type used in crosswords, it holds true.

For languages L made up from an English dictionary based on 26 letters plus \square , one may estimate the value of α_L to be around 0.6, in other words, there are no imminent shortage of crosswords. We shall return to this estimation in the last section of the paper.

Curiouser and curiouser!

We now approach higher dimensional crosswords. The size of the crosswords is then measured by n^d where n is the generalized notion of height or width and d is the *dimension* (which was 2 before). That is, we consider cubic (for $d = 3$) or even hyper-cubic (for $d > 3$) crosswords. The probability of one dimension (think: row) of the crosswords being valid equals

$$\left(\frac{c_L(n)}{|\mathcal{A}|^n} \right)^{n^{d-1}} = \frac{c_L(n)^{n^{d-1}}}{|\mathcal{A}|^{n^d}}.$$

This is almost the same result as before, but note the exponent n^{d-1} . In the case $d = 3$, where we might imagine the crossword as a cube made up of 'sticks' of sequences from L , the exponent corresponds to the fact that in each dimension there are n^2 sticks. The probability of all dimensions (think: rows and columns) being valid equals

$$\left(\frac{c_L(n)^{n^{d-1}}}{|\mathcal{A}|^{n^d}} \right)^d = \frac{c_L(n)^{dn^{d-1}}}{|\mathcal{A}|^{dn^d}}.$$

Again, this should come as no shock. The total number of possible crosswords (think: any matrix) is multiplied with the probability and we find:

$$N_n^{(d)} = |\mathcal{A}|^{n^d} \frac{c_L(n)^{dn^{d-1}}}{|\mathcal{A}|^{dn^d}} = \frac{c_L(n)^{dn^{d-1}}}{|\mathcal{A}|^{n^d(d-1)}}.$$

Applying logarithm yields

$$\log N_n^{(d)} = dn^{d-1} \log c_L(n) - n^d(d-1) \log |\mathcal{A}|$$

and reorganizing the terms,

$$\log N_n^{(d)} = \frac{dn^d}{\log |\mathcal{A}|} \left(\frac{\log_{|\mathcal{A}|} c_L(n)}{n} - \frac{d-1}{d} \right). \quad (2)$$

The leftmost fraction inside the parentheses above is recognized as α_n from before. We recall that the number α is used to denote the limiting value of the fraction as n becomes very big. We find, that if, say, $d = 3$ the value of α must be at least $\frac{2}{3}$ if we want to have many, big crosswords. As the dimension of the crosswords grow, the languages L must have larger and larger α -value to sustain the notion of many crosswords.

It seems like α_L expresses something fundamental about the language L , and indeed information theorists have a name for that value:

Definition 2 Let L be a language and set

$$\tilde{H}(L) = \lim_{n \rightarrow \infty} \frac{\log_{|\mathcal{A}|} c_L(n)}{n}. \quad (3)$$

whenever the limit exists. For languages where $\tilde{H}(L)$ is defined we say that $\tilde{H}(L)$ is the *entropy* of L .

We recognize the entropy as the same thing as we know as α_L . The symbol on top of \tilde{H} is there to remind us that this is a special kind of entropy: in traditional information theory entropy is defined a bit differently, and the quantity \tilde{H} is known as the topological entropy or Hartley entropy.

We may wonder what happens if, say, $\tilde{H}(L) = \frac{1}{4}$. What kind of crosswords are possible? Why, crosswords of dimension $d > \frac{4}{3}$ of course, since if $d = \frac{4}{3}$ we have $(d - 1)/d = \frac{1}{4}$. How to visualize a crossword in 1.333 dimensions is another matter, though, and probably better left as an exercise to the reader!

Languages can (and will often) have entropies that does not correspond to an integer dimension of crosswords. As an example, we now calculate the entropy of a simple language. Let A be the alphabet of three letters $\{a, b, c\}$ and let L be the language of all sequences made using just a and c . Thus $accac$ is in L but aba is not. The value of $c_L(n)$ is now seen to equal 2^n because there are exactly 2^n different sequences of length n made from a and c . We therefore find

$$\tilde{H}(L) = \lim_{n \rightarrow \infty} \frac{\log_3 2^n}{n} = \lim_{n \rightarrow \infty} \frac{n \log_3 2}{n} = \log_3 2 = \frac{\log 2}{\log 3}.$$

We used some facts about the logarithm function (namely that $\log 2^n = n \log 2$ and that $\log_3 x = \log x / \log 3$), but otherwise the calculation should be straightforward. Notice that the n disappear completely from the calculation, so we need not consider the existence of the limit.

A recapitulation

Let us briefly examine what we have learned so far: We introduced the concept of languages as being sets of finite sequences of letters. We have then made a satisfactory definition of what is a crossword over a language. Using elementary combinatorics and probability techniques we have estimated the number of valid crosswords of size $n \times n$ (or, in the case of other dimensions, size n^d). This number depends on the size of the alphabet, $|A|$, as well as the special function $c_L(n)$. We then observed, that there are essentially two different cases: In the first case ($\alpha_L < \frac{d-1}{d}$), the number of crosswords becomes very small as the size, n , grows. In the second case ($\alpha_L > \frac{d-1}{d}$), the same number becomes infinitely big as the size grows.

This calls for a reformulation of our initial question: While we opened the paper asking about the existence of crosswords, we are now tempted to ask: "Given a language L , what is the greatest dimension d for which there are many (big) crosswords?". This move encourages us to consider non-integer values of d , and thus we have definitely left the realm of ordinary crossword puzzles! The answer to the new question is related to $\tilde{H}(L)$ as we have just seen. In fact, combining the formulas (3) and (2), shows

$$\tilde{H}(L) = \frac{d_0 - 1}{d_0} \quad \text{and} \quad d_0 = \frac{1}{1 - \tilde{H}(L)},$$

where d_0 is exactly the largest dimension where it is possible to create (many) crosswords over L .

Note, that d_0 may be arbitrarily big, and it is reasonable to define d_0 to be infinite when $\tilde{H}(L) = 1$. How is that for a crossword puzzle! Actually, if $\tilde{H}(L) = 1$ it is indeed quite trivial to create crossword puzzles (in any dimension). An example of such a language L is that which is made up of every integer. The alphabet A is just the digits, and $c_L(n) = |A|^n = 10^n$ (because any sequence of length n made up from digits, is a valid number) so clearly $\tilde{H}(L) = 1$.

We have arrived at a concept of entropy by a quite unusual method. Aside from (hopefully) some expository advantages there are other reasons for picking this approach: We now have an entropy concept defined on the rather simple construction of a *language* or, at least, any language for which the limit of (3) exists. In contrast, the traditional entropy is based on the concept of information sources (that are also known as stochastic processes), that are in turn based on a quite technical probability theoretic framework.

This concludes the first section of the paper. The connection between the complexity of a language and the ability to create crosswords may not come as a surprise. But that this connection leads directly to entropy, the cornerstone of information theory is, at the very least, rather neat.

Notes and references

A few notes about the history of the results follows. The idea of linking crossword puzzles and entropy is, as mentioned in the beginning, quite old as it dates back to [Shannon, 1948]. The method of calculation that we have used above, is described briefly in the last part of [Immink et al., 1998] where a few bits of history on the result is also found. It turns out, that even though Shannon did not publish anything else on crosswords, the argument he had in mind when he wrote the passage in [Shannon, 1948], was something similar to the one given above.

The entropy \tilde{H} introduced in this part of the paper is related to the Hartley entropy and Hausdorff dimensions of ‘nice’ subsets of A^∞ (infinite sequences over A), when considered as subsets of $[0, 1]$. Or, if one considers arbitrary subsets of A^∞ , the entropy \tilde{H} might be interpreted as a form of the box counting dimension, see e.g. [Falconer, 1990]. The connection between entropy and Hausdorff dimension is described in [Billingsley, 1965] and interesting results in this direction can be found in [Ryabko, 1986].

II THE THEORY OF SELF SIMILAR SETS

Hausdorff measure and Hausdorff dimension

Let \mathbb{R}^n be the Euclidean n -space, and write $|x - y|$ for the distance between two points x and y from \mathbb{R}^n . Any subset $E \subseteq \mathbb{R}^n$ has a (possibly infinite) diameter, $\text{diam}(E)$, defined to be the supremum of the distance between any two points from E (we define $\text{diam}(\emptyset) = 0$).

Definition 3 Let $E \subseteq \mathbb{R}^n$ be any subset of \mathbb{R}^n . For any $\delta > 0$ a δ -cover of E is a countable collection, $\mathcal{C} = \{C_i\}_{i \in \mathbb{N}}$, of subsets of \mathbb{R}^n each of which has $\text{diam}(C_i) < \delta$ and such that $E \subseteq \bigcup_{i \in \mathbb{N}} C_i$.

Definition 4 Let $E \subseteq \mathbb{R}^n$ be any subset of \mathbb{R}^n and let $s \geq 0$ be a non-negative number. For any $\delta > 0$, let

$$\mathcal{H}_\delta^s(E) = \inf \sum_{i=1}^{\infty} \text{diam}(C_i)^s \quad (4)$$

where the infimum is taken over all δ -covers $\mathcal{C} = \{C_i\}_{i \in \mathbb{N}}$ of E . The s -dimensional Hausdorff measure is given by

$$\mathcal{H}^s(E) = \lim_{\delta \rightarrow 0} \mathcal{H}_\delta^s(E). \quad (5)$$

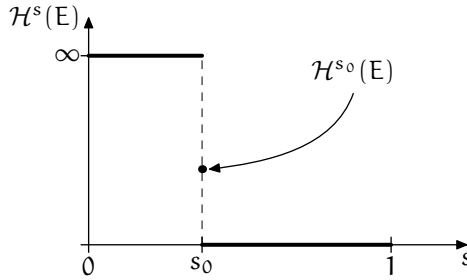
We must check that definition 4 is well made. It is seen immediately that to any $\delta > 0$ there is a δ -cover of E and thus the value of $\mathcal{H}_\delta^s(E)$ is well defined and non-negative (but it may be $+\infty$). It is also seen that $\mathcal{H}_\delta^s(E)$ is non-decreasing as $\delta \rightarrow 0$ since any δ -cover is also a δ' -cover provided $\delta \leq \delta'$ so that the infimum is taken over a decreasing number of covers, as δ becomes smaller. Thus the limit in (5) exists and the Hausdorff measure is well defined in $[0, \infty]$.

We can easily check that the Hausdorff measure is an outer measure on \mathbb{R}^n : The empty set is covered by the cover of empty sets so $\mathcal{H}^s(\emptyset) = 0$ for all values³ of s . Assuming $A \subseteq B$ any δ -cover of B is a δ -cover of A as well, hence $\mathcal{H}^s(A) \leq \mathcal{H}^s(B)$. For A and B subsets of \mathbb{R}^n and δ -covers for A and B respectively, a δ -cover of $A \cup B$ may be constructed by combining the sets covering A with those covering B . Consequently $\mathcal{H}^s(A \cup B) \leq \mathcal{H}^s(A) + \mathcal{H}^s(B)$ from which follows, by repeating the argument, the last requirement of an outer measure.

A characteristic property of the Hausdorff measure is the following. Fix a set $E \subseteq \mathbb{R}^n$ and consider the value of the Hausdorff measure for different values of s . Assume $s < t$, then the following calculation

$$\mathcal{H}_\delta^s(E) = \inf \sum \text{diam}(C_i)^s = \inf \sum \text{diam}(C_i)^{s-t+t} \leq \inf \sum \delta^{s-t} \text{diam}(C_i)^t = \delta^{s-t} \mathcal{H}_\delta^t(E)$$

shows that if $\mathcal{H}^s(E) > 0$ then $0 < \delta^{s-t} \mathcal{H}_\delta^t(E)$ for all $\delta > 0$ and hence $\mathcal{H}^t(E) = \infty$. Since $\mathcal{H}^s(E)$ is non-increasing with s , it follows that the Hausdorff measure has the following peculiar behavior:



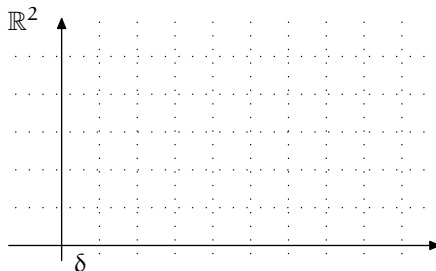
³We define $0^0 = 0$.

The spot where the Hausdorff measure “jumps” is well defined and given by $\inf\{s \mid \mathcal{H}^s(E) = 0\} = \sup\{s \mid \mathcal{H}^s(E) = \infty\}$.

Definition 5 The *Hausdorff dimension* of a set $E \subseteq \mathbb{R}^n$ is defined as

$$\dim_{\mathcal{H}}(E) = \inf\{s \mid \mathcal{H}^s(E) = 0\} = \sup\{s \mid \mathcal{H}^s(E) = \infty\}.$$

A variant of the ideas behind the Hausdorff leads to the Box-counting dimension, which we introduce now. Let $E \subseteq \mathbb{R}^n$ be any set. To any $\delta > 0$ we consider the δ -mesh of \mathbb{R}^n , that is, all n -cubes with edges of length equal to δ and aligned with the axis. In \mathbb{R}^2 the δ -mesh looks like this



The number of n -cubes from the δ -mesh that intersects E is denoted by $N_\delta(E)$ and the box counting dimensions may now be introduced.

Definition 6 For any non-empty and bounded set $E \subseteq \mathbb{R}^n$ the upper- and lower *box-counting dimension* of E are defined as

$$\overline{\dim}_{\mathcal{B}}(E) = \limsup_{\delta \rightarrow 0} \frac{\log N_\delta(E)}{-\log \delta} \quad (6)$$

and

$$\underline{\dim}_{\mathcal{B}}(E) = \liminf_{\delta \rightarrow 0} \frac{\log N_\delta(E)}{-\log \delta} \quad (7)$$

respectively. If $\overline{\dim}_{\mathcal{B}}(E) = \underline{\dim}_{\mathcal{B}}(E)$ we write $\dim_{\mathcal{B}}(E)$ for the common value which is known as the *box-counting dimension* of E .

The box-counting dimensions of a set E are quite easy to approximate simply by “counting” the number of small n -dimensional boxes (n -cubes, actually) that intersects with E . The smaller the size of the mesh, the more accurately the approximation will be. This procedure is the reason for the name of the box-counting dimension. It is possible to give other definitions of the box-counting dimension (an extensive coverage is given in [Falconer, 1990, section 3.1]). For example, let $N'_\delta(E)$ be the number of sets with diameter at most δ needed to cover E . The diameter of a n -cube with sides of unity is the length of the diagonal, that is, $\sqrt{1^2 + \dots + 1^2} = \sqrt{n}$. Therefore, the cubes from the δ/\sqrt{n} -mesh have diameter δ and hence $N'_\delta(E) \leq N_{\delta/\sqrt{n}}(E)$. We have then for any $\delta < 1$

$$\frac{\log N'_\delta(E)}{-\log \delta} \leq \frac{N_{\delta/\sqrt{n}}(E)}{-\log \delta} = \frac{N_{\delta/\sqrt{n}}(E)}{-\log(\delta/\sqrt{n}) - \log \sqrt{n}}$$

and taking limes supremum and limes infimum with $\delta \rightarrow 0$ yields

$$\limsup_{\delta \rightarrow 0} \frac{\log N'_\delta(E)}{-\log \delta} \leq \overline{\dim}_{\mathcal{B}}(E) \quad \text{and} \quad \liminf_{\delta \rightarrow 0} \frac{\log N'_\delta(E)}{-\log \delta} \leq \underline{\dim}_{\mathcal{B}}(E).$$

Conversely, if A is a set of diameter at most δ , then it can be contained inside 3^n cubes from the δ -mesh: Let C_1 be one cube that intersects A and let C_2, \dots, C_{3^n} be the neighbors of C_1 . Then to any point in $x \in A$ there is a point $y \in A \cap C_1$ with $|x - y| \leq \text{diam}(A) \leq \delta$, and thus $x \in \cup_{i=1}^{3^n} C_i$. It follows that $N_\delta(E) \leq 3^n N'_\delta(E)$ so that

$$\frac{\log N_\delta(E)}{-\log \delta} \leq \frac{n \log 3 + \log N'_\delta(E)}{-\log \delta}$$

and taking limes supremum and limes infimum as above, we find

$$\overline{\dim}_B(E) \leq \limsup_{\delta \rightarrow 0} \frac{n \log 3 + \log N'_\delta(E)}{-\log \delta} \quad \text{and} \quad \underline{\dim}_B(E) \leq \liminf_{\delta \rightarrow 0} \frac{n \log 3 + \log N'_\delta(E)}{-\log \delta}$$

which leads us to the conclusion that the box-counting dimensions may just as well be defined based on $N'(E)$. As a convenience we define $\overline{\dim}_B(\emptyset) = \underline{\dim}_B(\emptyset) = 0$ and $\overline{\dim}_B(\mathbb{R}^n) = \underline{\dim}_B(\mathbb{R}^n) = n$.

Lemma 1 For any set $E \subseteq \mathbb{R}^n$ we have $\dim_H(E) \leq \underline{\dim}_B(E) \leq \overline{\dim}_B(E)$.

Proof: The second inequality is trivial, by (6) and (7). To check the first inequality, observe that since a collection consisting of $N'_\delta(E)$ sets of diameter at most δ is a cover of E , then by (4) we have $\mathcal{H}_\delta^s(E) \leq N'_\delta(E) \delta^s$. Taking logarithm and rearranging on the right hand side, we get

$$\log \mathcal{H}_\delta^s(E) \leq s \log \delta + \log N'_\delta(E) = \log \delta \left(s - \frac{\log N'_\delta(E)}{-\log \delta} \right).$$

If $s > \liminf_{\delta \rightarrow 0} (\log N'_\delta(E)) / -\log \delta$ then $\liminf_{\delta \rightarrow 0} \log \mathcal{H}_\delta^s(E) \leq -\infty$ meaning $\mathcal{H}^s(E) = 0$ so that $\dim_H(E) \leq s$. Since this holds for any $s > \underline{\dim}_B(E)$ we have $\dim_H(E) \leq \underline{\dim}_B(E)$. \square

Iterated function schemes

Definition 7 A map $S : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a *contraction* if there exists a number $0 < c < 1$ such that

$$|S(x) - S(y)| \leq c|x - y| \tag{8}$$

for all $x, y \in \mathbb{R}^n$. If (8) holds with equality, we say that S is a *similarity*.

Definition 8 An *iterated function scheme* (IFS) is a finite set $S = \{S_1, \dots, S_k\}$ of contractions on $X \subseteq \mathbb{R}^n$, that is, maps $S_i : X \rightarrow X$ with

$$|S_i(x) - S_i(y)| \leq c_i|x - y| \tag{9}$$

for any $x, y \in X$ and suitable values $0 < c_i < 1$. To an IFS, S , we identify a map of sets $S : X \rightarrow X$ defined by $S(A) = \cup_{i=1}^k S_i(A)$ for any set $A \subseteq X$. A set $F \subseteq X$ is said to be *invariant* with respect to the IFS S , if

$$F = S(F) = \cup_{i=1}^k S_i(F). \tag{10}$$

Theorem 1 To any IFS $S = \{S_1, \dots, S_k\}$ on $X \subseteq \mathbb{R}^n$ there exists an unique non-empty compact set $F \subseteq X$ that is invariant with respect to S .

Before we move on to the proof of this theorem, we must introduce the *Hausdorff distance* between sets. We will use this to give an elegant proof of theorem 1, based on the fixed point theorem by Banach.

Definition 9 Let A be a subset of \mathbb{R}^n . For any $\delta > 0$ we write A_δ for the δ -neighborhood of A , that is, $A_\delta = \{x \in \mathbb{R}^n \mid \inf_{a \in A} d(x, a) \leq \delta\}$. For two subsets A and B of \mathbb{R}^n , the *Hausdorff distance* between the sets is defined as

$$d(A, B) = \inf\{\delta \mid A \subseteq B_\delta \text{ and } B \subseteq A_\delta\}.$$

Let $\mathcal{C} = \mathcal{C}(X)$ denote the non-empty compact subsets of X for $X \subseteq \mathbb{R}^n$. We will now check that d is in fact a metric on \mathcal{C} . Firstly, we have $d(A, B) \geq 0$ for any sets $A, B \in \mathcal{C}$ with equality if and only if $A = B$. The inequality follows from the definition of d (the value of d is an infimum of positive numbers) and the fact that for any $A \in \mathcal{C}$ the 0-neighborhood is the set itself, shows that $d(A, B) = 0$ when $A = B$. Assuming $d(A, B) = 0$ we have $A \subseteq B_\delta$ and $B \subseteq A_\delta$ for any $\delta > 0$. Hence to any point $a \in A$ we may select a sequence b_i with $|b_i - a| \leq 1/i$ and $b_i \in B$. By the compactness of B , it follows that the limit, that is a , is in B . By an equivalent argument it follows that $B \subseteq A$. Secondly, $d(A, B) = d(B, A)$ for any sets $A, B \in \mathcal{C}$. This follows trivially from the symmetry in the definition of d . Thirdly, assume $A, B, C \in \mathcal{C}$. For all $\delta_{AB} > d(A, B)$ and $\delta_{BC} > d(B, C)$ we have $A \subseteq B_{\delta_{AB}}$ and $B \subseteq C_{\delta_{BC}}$. But then $A \subseteq C_{\delta_{AB} + \delta_{BC}}$. Correspondingly we get $C \subseteq A_{\delta_{AB} + \delta_{BC}}$. Thus $d(A, C) \leq d(A, B) + d(B, C)$. The Hausdorff distance, d , is therefore also known as the *Hausdorff metric*.

Before we proceed, let us look at a few more easily derived properties of the Hausdorff metric. For any non-empty set $A \subset \mathbb{R}^n$ we have $d(A, \bar{A}) = 0$ where \bar{A} is the closure of A . To see this, observe that to any $\delta > 0$ and any $x \in \bar{A}$ we have $B(x, \delta) \cap A \neq \emptyset$. This shows that $\bar{A} \subseteq A_\delta$. Because $A \subseteq \bar{A} \subseteq \bar{A}_\delta$ holds trivially, we have $d(A, \bar{A}) \leq \delta$, and since δ was chosen freely, we conclude

$$d(A, \bar{A}) = 0. \tag{11}$$

Assume now that $(A_i)_{i \in \mathbb{N}}$ and $(B_i)_{i \in \mathbb{N}}$ are families of subsets of \mathbb{R}^n . We then argue that

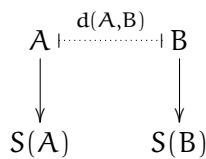
$$d\left(\bigcup_{i=1}^{\infty} A_i, \bigcup_{i=1}^{\infty} B_i\right) \leq \sup_{i \in \mathbb{N}} d(A_i, B_i). \tag{12}$$

To see this, let $\delta = \sup_{i \in \mathbb{N}} d(A_i, B_i)$ and assume $x \in \bigcup_{i=1}^{\infty} A_i$. In particular, $x \in A_j$ for some value of j , and hence we have $x \in (B_j)_{d(A_j, B_j)} \subseteq (B_j)_\delta$. We conclude that $\bigcup_{i=1}^{\infty} A_i \subseteq (\bigcup_{i=1}^{\infty} B_i)_\delta$, and the symmetric inclusion follows likewise. From this follows (12), and as a trivial consequence we have $d(A, \bigcup_{i=1}^{\infty} B_i) \leq \sup_{i \in \mathbb{N}} d(A, B_i)$.

The final property of the Hausdorff metric we will mention, is the following: Let A and B be subsets of \mathbb{R}^n and let $S : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a contraction with scaling constant $0 < c < 1$. Then

$$d(S(A), S(B)) \leq cd(A, B). \tag{13}$$

To see that (13) holds, we must first check that $S(A) \subseteq (S(B))_{cd(A, B)}$. The following simple diagram may be of help in the next few calculations⁴.



⁴... which may be thought of as a variant of diagram chasing.

For $x \in S(A)$ we have $x' \in A$ with $S(x') = x$. We have $A \subseteq B_{d(A,B)}$ so there exists an element $y' \in B$ with $|x' - y'| \leq d(A, B)$. Setting $y = S(y')$ we find, by the contraction property of S , that $|S(x') - S(y')| = |x - y| \leq c|x' - y'| \leq cd(A, B)$, and hence $x \in (S(B))_{cd(A,B)}$. The symmetric inclusion $S(B) \subseteq (S(A))_{cd(A,B)}$ follows by the same method and thus (13) has been established.

Theorem 2 Let $\mathcal{C} = \mathcal{C}(X)$ denote the non-empty compact subsets of X with $X \subseteq \mathbb{R}^n$ and let $d : \mathcal{C} \times \mathcal{C} \rightarrow [0, \infty[$ be the Hausdorff metric. Then (\mathcal{C}, d) is a complete metric space.

Proof: Assume $(A_i)_{i \in \mathbb{N}}$ is a Cauchy sequence of sets in (\mathcal{C}, d) . We may further assume that $\bigcup_{i=1}^{\infty} A_i$ is bounded since if it is not, we can replace $(A_i)_{i \in \mathbb{N}}$ with its own tail, namely $(A_i)_{i > N}$ where N is a number chosen so that $d(A_n, A_m) \leq 1$ whenever $n, m > N$. Define $E_j = \overline{\bigcup_{i=j}^{\infty} A_i}$ and observe that

$$E_1 \supseteq E_2 \supseteq \dots, \quad (14)$$

The non-increasing sequence $(E_j)_{j \in \mathbb{N}}$ consists of compact sets so its intersection

$$E = \bigcap_{j=1}^{\infty} E_j$$

is again a compact set. We now use the Cauchy property to show that in fact A_i converges to E in the Hausdorff metric. Let $\delta > 0$ be given and pick accordingly a number $j \in \mathbb{N}$ such that $d(A_n, A_m) \leq \delta$ provided $n, m > j$. We first seek the inequality $d(A_j, E_j) \leq \delta$. But this follows by

$$\begin{aligned} d(A_j, E_j) &\leq d(A_j, \bigcup_{i=j}^{\infty} A_i) + d(\bigcup_{i=j}^{\infty} A_i, \overline{\bigcup_{i=j}^{\infty} A_i}) \\ &\stackrel{(11)}{=} d(A_j, \bigcup_{i=j}^{\infty} A_i) \\ &\stackrel{(12)}{\leq} \sup_{i \geq j} d(A_j, A_i) \leq \delta \end{aligned}$$

where the first inequality by d being a metric.

We then seek the inequality $d(E_j, E) \leq \delta$. It is clear that $E \subseteq E_j \subseteq (E_j)_{\delta}$. Assume therefore $x \in E_j$ and it follows, that $x \in A_i$ for some $i \geq j$. For $k \leq i$ we have by (14) that $x \in E_k$. For any $k > i$ we have $d(A_i, A_k) \leq \delta$ so that $x \in (A_k)_{\delta}$ and hence also $x \in (E_k)_{\delta}$. For any $k > i$ we thus choose x_k from E_k such that $|x_k - x| \leq \delta$. Because $x_k \in E$ and E is compact, there will be a convergent subsequence of x_k the limit of which lies in E . Hence $x \in E_{\delta}$. We conclude that $d(E_j, E) \leq \delta$.

Finally, we find that $d(A_j, E) \leq d(A_j, E_j) + d(E_j, E) \leq 2\delta$ so that A_j converges to E in the Hausdorff metric. \square

Proof: (of theorem 1) Observe first, that the map S is a contraction on (\mathcal{C}, d) . To any $A, B \in \mathcal{C}$ we have

$$\begin{aligned} d(S(A), S(B)) &= d\left(\bigcup_{i=1}^k S_i(A), \bigcup_{i=1}^k S_i(B)\right) \\ &\stackrel{(12)}{\leq} \max_{i=1, \dots, k} d(S_i(A), S_i(B)) \\ &\stackrel{(13)}{\leq} \max_{i=1, \dots, k} c_i d(A, B) = cd(A, B) \end{aligned}$$

for a suitable constant $0 < c < 1$. By Banach's fixed point theorem, there is a unique fixed point to any contraction on a complete metric space. As we have just seen, S is a contraction on the complete metric space (\mathcal{C}, d) , so we conclude that there is a unique non-empty compact set $F \in \mathcal{C}$ such that

$$F = S(F),$$

and that for any non-empty compact set $F_1 \in \mathcal{C}$ we get, by setting $F_2 = S(F_1)$ and $F_i = S(F_{i-1})$, a sequence of sets F_1, F_2, \dots convergent with F as limit. \square

As an important special case, we find from the above proof that if F_1 is a non-empty compact set with $S(F_1) \subseteq F_1$, then $S(S(F_1)) \subseteq S(F_1)$ and we have a *non increasing* sequence

$$F_1 \supseteq F_2 \supseteq \dots$$

of sets, the limit of which is then the intersection, $\bigcap_{i=1}^{\infty} F_i$. The following example shows how this can be used to give a good visual representation of what the limit set looks like.

Example 1 The IFS given by $S = \{S_1, S_2\}$, where the functions are defined on $[0, 1]$ in the following way,

$$\begin{aligned} S_1(x) &= \frac{1}{3}x \\ S_2(x) &= \frac{1}{3}x + \frac{2}{3} \end{aligned} \tag{15}$$

induces the well known Cantor set as its invariant set. Recall that the Cantor set is obtained by removing, iteratively, the middle thirds of all remaining line segments starting with $[0, 1]$ (from which $]\frac{1}{3}, \frac{2}{3}[$ is then removed etc.). Writing $C_1 = [0, 1]$, $C_2 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$ and so forth, we find that $C_1 \supseteq C_2 \supseteq \dots$ and hence $C = \bigcap_{i=1}^{\infty} C_i = \lim_{i \rightarrow \infty} C_i$. It is best to think of this construction visually,

$$\begin{array}{l} \text{—————} C_1 \\ \text{———} \quad \quad \quad \text{———} C_2 \\ \text{——} \quad \quad \quad \text{——} \quad \quad \quad \text{——} \quad \quad \quad \text{——} C_3 \\ \text{—} \quad \text{—} \quad \quad \text{—} \quad \text{—} \quad \quad \text{—} \quad \text{—} \quad \quad \text{—} \quad \text{—} C_4 \\ \text{..} \quad \text{..} \quad \quad \text{..} \quad \text{..} \quad \quad \text{..} \quad \text{..} \quad \quad \text{..} \quad \text{..} C_5 \\ \text{...} \quad \text{...} \quad \quad \text{...} \quad \text{...} \quad \quad \text{...} \quad \text{...} \quad \quad \text{...} \quad \text{...} C_6 \\ \text{...} \quad \text{...} \quad \quad \text{...} \quad \text{...} \quad \quad \text{...} \quad \text{...} \quad \quad \text{...} \quad \text{...} C_7 \end{array} \tag{16}$$

where the sets C_i are shown for small values of i . The sets C_i are known as *pre-sets* of the Cantor set C , for obvious reasons⁵. To see that C is invariant with respect to S , we must check that $C = S_1(C) \cup S_2(C)$. We clearly have $S_1(C_i) \cup S_2(C_i) = C_{i+1}$ and it follows that

$$\lim_{i \rightarrow \infty} C_i = \lim_{i \rightarrow \infty} S_1(C_i) \cup \lim_{i \rightarrow \infty} S_2(C_i)$$

from which we get $C = S_1(C) \cup S_2(C)$. According to theorem 1 the Cantor set is the only non-empty compact set that is invariant with respect to S . In fact, because $C_1 = [0, 1]$ is a non-empty compact set, we did not even need to check for C being invariant; it must be so by the fixed point theorem of Banach, as we argued in the proof of theorem 1.

⁵The Cantor set is part of a loosely defined class of sets known as *fractals*. Many fractals, including as we have just seen the Cantor set, can be constructed by repeated removal of subsets from a larger set, and the sequence of sets produced by such a procedure are known as *pre-fractals*.

Hausdorff dimension of self-similar sets

If all the contractions of an IFS are in fact similarities, we say that the corresponding invariant set is a *self-similar* set. It turns out, that the Hausdorff dimension of such sets, can be calculated quite easily.

Lemma 2 Let c_1, \dots, c_k be numbers with $0 < c_i < 1$ for $i = 1, \dots, k$. There exists a unique number s such that

$$\sum_{i=1}^k c_i^s = 1. \quad (17)$$

Proof: Let c_1, \dots, c_k be given, and write $\alpha(s)$ for the sum (17). It is now seen that $\alpha(s) \searrow 0$ as $s \rightarrow \infty$. However, $\alpha(0) = k > 1$ and since α is continuous, there exists a unique $s > 0$ for which $\alpha(s) = 1$. \square

Definition 10 Let $S = \{S_1, \dots, S_k\}$ be an IFS of similarities with corresponding constants $0 < c_i < 1$ for $i = 1, \dots, k$. The (according to lemma 2) uniquely defined number s such that $\sum_{i=1}^k c_i^s = 1$ is known as the *similarity dimension* of S .

Example 2 For the similarities S_1 and S_2 defining the Cantor set (see example 1), we have $c_1 = c_2 = \frac{1}{3}$ and hence (17) takes the form $(\frac{1}{3})^s + (\frac{1}{3})^s = 1$ which may be rearranged to get $(\frac{1}{3})^s = \frac{1}{2}$. Taking logarithm on both sides, we then find that $s = (\log 2)/(\log 3)$ is the similarity dimension of the Cantor set.

In the following we employ the notion of *outer measures* on subsets of \mathbb{R}^n . Recall, that an outer measure on $X \subseteq \mathbb{R}^n$ is a map μ which to any subset $A \subseteq X$ assigns a non-negative (but maybe infinite) number. Furthermore, the following properties must hold: Firstly, $\mu(\emptyset) = 0$; secondly, $\mu(A) \leq \mu(B)$ if $A \subseteq B \subseteq X$; and thirdly, for any countable sequence of sets $(A_i)_{i \in \mathbb{N}}$ with $A_i \subseteq X$, we have $\mu(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$.

The following lemma is a valuable tool for getting lower bounds on Hausdorff dimensions of sets.

Lemma 3 (MASS DISTRIBUTION PRINCIPLE) Let $E \subseteq \mathbb{R}^n$ be a set and let μ be a outer measure \mathbb{R}^n with support on E and with bounded yet non-trivial total measure, $0 < \mu(E) < \infty$. Assume there are numbers $s \geq 0$, $\epsilon > 0$ and $c > 0$ such that for any set $C \in \mathbb{R}^n$ with $\text{diam}(C) < \epsilon$ we have $\mu(C) \leq c \text{diam}(C)^s$. Then the s -dimensional Hausdorff measure of E is bounded by $\mathcal{H}^s(E) > \mu(E)/c$ and the Hausdorff dimension is bounded by $\dim_{\text{H}}(E) \geq s$.

Proof: Recall (from definitions 3 and 4) that the the s -dimensional Hausdorff measure $\mathcal{H}^s(E)$, is defined as $\lim_{\delta \rightarrow 0} \mathcal{H}_{\delta}^s(E)$, with $\mathcal{H}_{\delta}^s(E) = \inf_{\mathcal{C}} \sum_{C \in \mathcal{C}} \text{diam}(C)^s$ and the infimum taken over all δ -covers \mathcal{C} of E . For any $\delta < \epsilon$ and any δ -cover \mathcal{C} of E , we have

$$0 < \mu(E) = \mu\left(\bigcup_{C \in \mathcal{C}} C\right) \leq \sum_{C \in \mathcal{C}} \mu(C) \leq \sum_{C \in \mathcal{C}} c \text{diam}(C)^s.$$

Taking the infimum over all δ -covers it follows that $\mathcal{H}_{\delta}^s(E) \geq \mu(E)/c$. Letting δ approach zero, we find that $\mathcal{H}^s \geq \mu(E)/c$ and we can conclude that $\dim_{\text{H}}(E) \geq s$. \square

A frequently used method of constructing outer measures supported on self-similar sets, is to use the mass distribution principle with repeated subdivision. The following lemma shows how.

Lemma 4 Let there be given a k -ary subdivision of $X \subseteq \mathbb{R}^n$ in the following way: For any $t \in \mathbb{N}$ there are non-empty compact subsets X_{i_1, \dots, i_t} of X indexed by vectors of the form $(i_1, \dots, i_t) \in \{1, \dots, k\}^t$ and such that

$$X_{i_1, \dots, i_t} \supseteq \bigcup_{j=1}^k X_{i_1, \dots, i_t, j}. \quad (18)$$

Write \mathcal{X} for all sets of the form X_{i_1, \dots, i_t} including the set X . Assume further, that a map m is given which assigns non-negative finite mass to all sets from \mathcal{X} in such a way that

$$m(X_{i_1, \dots, i_t}) = \sum_{j=1}^k m(X_{i_1, \dots, i_t, j}). \quad (19)$$

Let $E = \bigcap_t \bigcup_{(i_1, \dots, i_t)} X_{i_1, \dots, i_t}$. Assume $\text{diam}(X_{i_1, \dots, i_t}) \rightarrow 0$ and $m(X_{i_1, \dots, i_t}) \rightarrow 0$ both for $t \rightarrow \infty$ and regardless of the actual values of i_1, i_2, \dots . Then m may be extended to an outer measure μ on E by the definition

$$\mu(A) = \inf \sum_{C \in \mathcal{C}} m(C) \quad (20)$$

where the infimum is taken over all collections of sets $\mathcal{C} \subseteq \mathcal{X}$ where $A \cap E \subseteq \bigcup_{C \in \mathcal{C}} C$.

Proof: We begin by establishing that E is well defined. Set $E_t = \bigcup_{(i_1, \dots, i_t)} X_{i_1, \dots, i_t}$ and observe that being a union of a finite number of compact sets, E_t is itself compact. By (18) we see that $(E_t)_{t \in \mathbb{N}}$ is a decreasing sequence of non-empty compact sets and so E is well defined and not empty.

Assume A is any subset of X . Taking $\mathcal{C} = \{X\}$ shows that $\mu(A)$ is well defined and less than or equal to $m(X)$, and hence finite. We easily find $\mu(\emptyset) = 0$ simply by letting \mathcal{C} be empty. Summing over an empty set then yields 0 by definition. Assume then $A \subseteq B \subseteq X$. If $\bigcup_{C \in \mathcal{C}} C$ covers $B \cap E$ then it clearly covers $A \cap E$ as well and hence $\mu(A) \leq \mu(B)$. Finally, let A and B be subsets of X . If $\bigcup_{C \in \mathcal{C}} C$ covers $A \cap E$ and $\bigcup_{D \in \mathcal{D}} D$ covers $B \cap E$, then $\mathcal{C} \cup \mathcal{D}$ contains sets covering $(A \cup B) \cap E$ and hence $\mu(A \cup B) \leq \mu(A) + \mu(B)$. Repeating the last argument inductively, yields the sub-additive property of an outer measure.

Let A be any open subset of X with empty intersection with E . It is then clear from (20) that $\mu(A) = 0$, and hence μ has support inside E . \square

Example 3 Let us try and employ the Mass distribution principle together with the above mentioned method of construction an outer measure to get a lower bound on the Hausdorff dimension of the Cantor set. The construction of the Cantor set by repeated removal of the “middle third” lends itself to the construction of an outer measure. (It might be helpful to keep (16) in mind through the following passage.) The pre-sets of the Cantor set can be perceived as a repeated subdivision of the unit interval. Using the notation from above, we have $X = [0, 1]$, then $X_1 = [0, \frac{1}{3}]$, $X_2 = [\frac{2}{3}, 1]$, then $X_{1,1} = [0, \frac{1}{9}]$, $X_{1,2} = [\frac{2}{9}, \frac{3}{9}]$, and so forth. Our mass distribution is now introduced as $m(X) = 1$, then $m(X_1) = \frac{1}{2}$, $m(X_2) = \frac{1}{2}$, then $m(X_{1,1}) = \dots = m(X_{2,2}) = \frac{1}{4}$ and so on. In general, we see that the intervals of length 3^{-t} are assigned a mass of 2^{-t} . We therefore have an outer measure on the unit interval which agrees with the mass function m on all sets of the form X_{i_1, \dots, i_t} . Assume now that $C \subseteq [0, 1]$ and select t so that $3^{-t-1} \leq \text{diam}(C) < 3^{-t}$. Since μ has support on the Cantor set we have $\mu(C) = \mu(C \cap E)$, where E is the Cantor set, it is now seen, that $C \cap E$ is covered by a single set

of the form X_{i_1, \dots, i_t} . Thus

$$\mu(C) = \mu(C \cap E) \leq 2^{-t} = 3^{-t(\log 2/\log 3)} \leq (3 \operatorname{diam}(C))^{\log 2/\log 3},$$

and hence the Mass distribution principle applies with $s = \log 2/\log 3$ and $c = 3$ and we have $\dim_{\text{H}}(E) \geq \log 2/\log 3$.

The technique used in the previous example will be used to prove a much more general result which holds for all self-similar sets. However, not all self-similar sets are as nice as the Cantor set. In particular we need to focus on the separating properties of the similarities defining the set.

Definition 11 Let $S = \{S_1, \dots, S_k\}$ be an IFS and let E be the unique non-empty compact invariant set corresponding to S . If there exists a non-empty open and bounded set G such that $S_i(G) \cap S_j(G) = \emptyset$ for all $i \neq j$ and such that

$$S(G) \subseteq G, \tag{21}$$

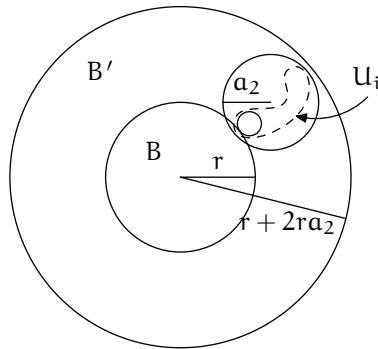
we say that S satisfies the *open set condition*.

Example 4 The separation condition of definition 11 is clearly fulfilled by the IFS defining the Cantor set, since we have $S_1([0, 1]) \cap S_2([0, 1]) = \emptyset$. Taking G to be the interior of $[0, 1]$ we may check the open set condition explicitly: The two similarities map G to $]0, 1/3[$ and $]2/3, 1[$ respectively, and these two sets are disjoint while their union is contained in $G =]0, 1[$.

It is worth noting, that the closure of G is a non-empty compact subset of \mathbb{R}^n and therefore we have $\lim_{i \rightarrow \infty} d(S^i(\overline{G}), E) = 0$ by the Banach fixed point theorem. Since $\overline{G} \supset S(\overline{G}) \supset S^2(\overline{G}) \dots$ is a decreasing sequence of non-empty compact sets, there is a non-empty compact intersection $\bigcap_{i=1}^{\infty} S^i(\overline{G})$ which equals E .

Lemma 5 Let $\{U_i\}$ be a disjoint family of open sets. Assume that there are numbers $r > 0$, $a_1 > 0$ and $a_2 > a_1$ such that each of the sets U_i contains a ball of radius ra_1 and is contained in a ball of radius ra_2 , then any ball, B , of radius r intersects with at most $(1 + 2a_2)^n (a_1)^{-n}$ of the sets $\{\overline{U}_i\}$.

Proof: Let B be a ball of radius r and assume $\overline{U}_1, \dots, \overline{U}_t$ all intersects with B . Enlarging the radius of B to yield another ball, B' , with radius $r + 2ra_2 = (1 + 2a_2)r$ ensures that all $\overline{U}_1, \dots, \overline{U}_t$ are contained in B' . The following drawing visualize this argument:



However, the sets U_1, \dots, U_t are disjoint and each contain a ball of volume $\nu_n(ra_2)^n$, with ν_n a suitable constant. The sets are all contained in the ball B' of radius $(1 + 2a_2)r$, so we have

$$t\nu_n(ra_2)^n \leq \nu_n((1 + 2a_2)r)^n.$$

This leads to $t \leq (1 + 2a_2)^n(a_2)^{-n}$. □

Theorem 3 (DIMENSION FORMULA OF SELF-SIMILAR SETS) Let $S = \{S_1, \dots, S_k\}$ be an IFS of similarities and assume S satisfies the open set condition. Let E be the invariant for S . Then the Hausdorff dimension of E equals the similarity dimension of S . In addition $\dim_H(E) = \dim_B(E)$.

Proof: We prove the theorem by showing that $0 < \mathcal{H}^s(E) < \infty$. The hard part of this is to see that $0 < \mathcal{H}^s(E)$, and so we begin by attacking this problem. The idea is to use a mass distribution principle similar to lemma 3.

Observe first, that by the open set condition (see (21)) we have a bounded non-empty open set G such that $G \supseteq S(G)$, and since \bar{G} is compact, the sequence of sets $\bar{G} \supseteq S(\bar{G}) \supseteq S^2(\bar{G}) \supseteq \dots$ is decreasing and convergent (in the Hausdorff metric) towards E . Furthermore, the sets $S_i(G)$ for $i = 1, \dots, k$ are disjoint, and hence so are the sets $S_i(S_j(G))$ for any combination of i, j and so forth. To formalize these ideas, writing S_{i_1, \dots, i_t} for $S_{i_1} \circ \dots \circ S_{i_t}$ we have $S^t(A) = \bigcup_{(i_1, \dots, i_t)} S_{i_1, \dots, i_t}(A)$ and in the case $S^t(G) = \bigcup_{(i_1, \dots, i_t)} S_{i_1, \dots, i_t}(G)$, the union is disjoint.

Let s be the similarity dimension of S , and let c_i be the similarity constant corresponding to S_i . We distribute a mass of 1 onto the sets of the form $S_{i_1, \dots, i_t}(G)$ by the following procedure: We assign the total mass of 1 to the set G , writing $m(G) = 1$. We then re-distribute the total mass 1 according to the following rule: For each $i = 1, \dots, k$ the set $S_i(G)$ is assigned mass $c_i^s m(G) = c_i^s$. Repeating the same construction yields an assignment of $(c_i c_j)^s$ to the set $S_{i,j}(G)$ for any $i, j \in \{1, \dots, k\}$, and we may continue this iteratively yielding a distribution of mass onto all sets of the form $S_{i_1, \dots, i_t}(G)$. By the usual construction we get an outer measure with support in $\bigcap_{t=1}^{\infty} \bigcup_{(i_1, \dots, i_t)} S_{i_1, \dots, i_t}(\bar{G}) = \bigcap_{t=1}^{\infty} S^t(\bar{G}) = E$. The measure is given by

$$\mu(A) = \inf_{\mathcal{C}} \sum_{C \in \mathcal{C}} m(C)$$

where \mathcal{C} is any family of sets of the form $S_{i_1, \dots, i_t}(\bar{G})$ and such that $A \subseteq \bigcup_{C \in \mathcal{C}} C$.

We are now ready to employ the mass distribution principle. Let A be any subset of \mathbb{R}^n with $\text{diam}(A) < 1$. Then $A \subseteq B$ where B is a ball of radius $r = \text{diam}(A) < 1$; we will estimate the measure of B . Define

$$\mathcal{X} = \{S_{i_1, \dots, i_t}(G) \mid rc_{\min} \leq c_{i_1} \cdots c_{i_t} \leq r < c_{i_1} \cdots c_{i_{t-1}}\}$$

where $c_{\min} = \min\{c_1, \dots, c_k\}$. We might say, that \mathcal{X} is a collection of images of G , obtained by applying a sequence of the similarities to G until the diameter of the image is just below r . Since G was the set from the open set condition (see definition 11), the union $\bigcup_{i=1}^k S_i(G)$ is disjoint. It follows that $\bigcup_{j=1}^k S_{i,j}(G)$ is disjoint for all $i = 1, \dots, k$, and in general $\bigcup_{(i_1, \dots, i_t)} S_{i_1, \dots, i_t}(G)$ is disjoint. From this we deduce that the sets of \mathcal{X} are disjoint, for if $S_{i_1, \dots, i_t}(G)$ and $S_{j_1, \dots, j_s}(G)$ are sets from \mathcal{X} with $t \leq s$ and if $S_{i_1, \dots, i_t} \cap S_{j_1, \dots, j_s}$ is non-empty, then we have $i_1 = j_1$ since $S_i(G) \cap S_j(G) = \emptyset$ for $i \neq j$. Repeating this argument shows that $i_m = j_m$ for $m = 1, \dots, t$. It remains to be checked if $t < s$ can happen, but if $S_{i_1, \dots, i_t} \in \mathcal{X}$ then no set of the form $S_{i_1, \dots, i_t, j}$ can be in \mathcal{X} , so we conclude $s = t$.

We use lemma 5 in the following. Let a_1 and $a_2 > a_1$ be given such that G is contained in a ball of radius a_2 and such that G contains a ball of radius a_1 . Then each $S_{i_1, \dots, i_t}(G) \in \mathcal{X}$ is contained in a ball of radius $c_{i_1} \cdots c_{i_t} a_2 \leq r a_2$ and contains a ball of radius $c_{i_1} \cdots c_{i_t} a_1 \geq c_{\min} r a_1$. It follows, by lemma 5, that B intersects with at most $(1 + 2r a_2)^n (c_{\min} r a_1)^{-n}$ of the sets $\{S_{i_1, \dots, i_t}(\bar{G}) \mid S_{i_1, \dots, i_t}(G) \in \mathcal{X}\}$. This fact enables us to estimate the measure of B . Recall that

$$\mu(B) = \inf \left\{ \sum_{C \in \mathcal{C}} m(C) \mid B \cap E \subseteq \cup_{C \in \mathcal{C}} C \right\}$$

and let $\mathcal{Y} = \{S_{i_1, \dots, i_t}(G) \mid S_{i_1, \dots, i_t}(G) \in \mathcal{X} \text{ and } B \cap S_{i_1, \dots, i_t}(\bar{G}) \neq \emptyset\}$. Then we have $B \cap E \subseteq \cup_{C \in \mathcal{Y}} C$ and hence

$$\begin{aligned} \mu(B) &\leq \sum_{S_{i_1, \dots, i_t}(G) \in \mathcal{Y}} m(S_{i_1, \dots, i_t}(G)) \\ &= \sum_{S_{i_1, \dots, i_t}(G) \in \mathcal{Y}} (c_{i_1} \cdots c_{i_t})^s \\ &\leq |\mathcal{Y}| r^s \\ &\leq (1 + 2a_2)^n (c_{\min} a_1)^{-n} r^s \end{aligned}$$

Setting $c = (1 + 2a_2)^n (c_{\min} a_1)^{-n}$ we get $\mu(B) \leq c \text{diam}(A)^s$ and the mass distribution principle now yields $\dim_{\text{H}}(E) \geq s$.

Recall that \mathcal{X} is a disjoint collection of sets of the form $S_{i_1, \dots, i_t}(G)$ such that $c_{i_1} \cdots c_{i_t} \leq r$ for some $r < 1$. Since for any t we have $E \subseteq \cup_{(i_1, \dots, i_t)} S_{i_1, \dots, i_t}(\bar{G})$ the set $\{S_{i_1, \dots, i_t}(\bar{G}) \mid S_{i_1, \dots, i_t}(G) \in \mathcal{X}\}$ is a cover of E . Let s be the similarity dimension of E , then

$$\sum_{S_{i_1, \dots, i_t} \in \mathcal{X}} (c_{i_1} \cdots c_{i_t})^s = 1$$

and it follows by the definition of \mathcal{X} that $|\mathcal{X}| c_{\min}^s r^s \leq 1$ leading to $|\mathcal{X}| \leq (c_{\min} r)^{-s}$. In other words, we have a cover of E made from sets of the form $S_{i_1, \dots, i_t}(\bar{G})$ each of with a diameter that may be bounded by

$$\text{diam}(S_{i_1, \dots, i_t}(\bar{G})) = c_{i_1} \cdots c_{i_t} \text{diam}(\bar{G}) \leq r \text{diam}(\bar{G}), \quad (22)$$

and the number of sets in the cover is at most $(c_{\min} r)^{-s}$. In terms of box-counting dimensions, we have $N'_r(E) \leq (c_{\min} r)^{-s}$ so that

$$\limsup_{r \rightarrow 0} \frac{\log N'_r(E)}{-\log r} \leq \limsup_{r \rightarrow 0} \frac{-s(\log r + \log c_{\min})}{-\log r} = s$$

and hence $\overline{\dim}_{\text{B}}(E) \leq s$. But by lemma 1 this implies $s \leq \dim_{\text{H}}(E) \leq \overline{\dim}_{\text{B}}(E) \leq s$ and hence $\dim_{\text{H}}(E) = \overline{\dim}_{\text{B}}(E) = s$. □

Theorem 4 Let $S = \{S_1, \dots, S_k\}$ be an IFS of contractions satisfying the open set condition and with corresponding constants $0 < c_i < 1$ for $i = 1, \dots, k$. Let E denote the invariant set with respect to S and let s be the number for which $\sum_{i=1}^k c_i^s = 1$. Then $\dim_{\text{H}} E \leq s$ and $\overline{\dim}_{\text{B}} E \leq s$.

Proof: The last paragraph of the proof of theorem 3 may be used by changing only the equal sign in (22) to a less than or equal sign. □

Theorem 5 Let $S = \{S_1, \dots, S_k\}$ be an IFS of contractions and assume that

$$b_i |x - y| \leq |S_i(x) - S_i(y)|$$

for all $x, y \in \mathbb{R}^n$ and with $0 < b_i < 1$. Let F be the invariant set with respect to S . Then $\dim_{\text{H}} F \geq s$ where s is the number for which $\sum_{i=1}^k b_i^s = 1$.

Proof: See e.g. [Falconer, 1990, Proposition 9.7]. □

Example 5 Let $S = \{S_1, S_2\}$ be an IFS which lives on $[0, 1]$ and with $S_1(x) = x/(2 + x)$ and $S_2(x) = 2/(2 + x)$. The invariant set, E , corresponding to S may be visualized by the following “first iterations”:



It is clear that the maps S_1 and S_2 are contractions. We now compute their respective contraction constants. We may use the mean value theorem which says that $|S_1(x) - S_1(y)| = S'_1(\alpha)|x - y|$ for some value α between x and y . In particular, we have

$$\left(\inf_{z \in [x, y]} |S'_1(z)| \right) |x - y| \leq |S_1(x) - S_1(y)| \leq \left(\sup_{z \in [x, y]} |S'_1(z)| \right) |x - y|,$$

where $[x, y]$ is denotes the numbers between x and y . The derivatives of the contraction maps are $S'_1(x) = 2/(x^2 + 4x + 4)$ and $S'_2(x) = -2/(x^2 + 4x + 4)$ and they are monotone in $[0, 1]$, so $S'_1(0) = -S'_2(0) = 0.5$ and $S'_1(1) = -S'_2(1) = 2/9$ are the important values. It follows that $2/9|x - y| \leq |S_i(x) - S_i(y)| \leq 0.5|x - y|$ for $i = 1, 2$. Using theorem 4 and 5 we have $t \leq \dim_{\text{H}}(E) \leq s$ where s solves $2(1/2)^s = 1$ and t solves $2(2/9)^t = 1$. The solutions are $s = 1$ and $t = 0.46$ and it is clear the the upper bound $s = 1$ on the Hausdorff dimension is useless since we know that $E \subset \mathbb{R}$. We therefore use another approach which is to observe that E may also be defined by the IFS $S = \{S_{1,1}, S_{1,2}, S_{2,1}, S_{2,2}\}$ where $S_{i,j} = S_i \circ S_j$. A few calculations yield

$$S_{1,1}(x) = \frac{x}{3x + 4}, \quad S_{2,1}(x) = \frac{2x + 4}{3x + 4}$$

$$S_{1,2}(x) = \frac{1}{x + 3}, \quad S_{2,2}(x) = \frac{x + 2}{x + 3}$$

and we find the derivatives of these to be

$$S'_{1,1}(x) = -S'_{2,1}(x) = \frac{4}{(3x + 4)^2}$$

and

$$S'_{1,2}(x) = -S'_{2,2}(x) = \frac{-1}{(x + 3)^2}.$$

We now set $b_{1,1} = b_{2,1} = \inf_{x \in [0,1]} |S'_{1,1}(x)| = 0.082$ and $c_{1,1} = c_{1,2} = \sup_{x \in [0,1]} |S'_{1,1}(x)| = 0.25$. Also, set $b_{1,2} = b_{2,2} = \inf_{x \in [0,1]} |S'_{1,2}(x)| = 0.0625$ and $c_{1,2} = c_{2,2} = \sup_{x \in [0,1]} |S'_{1,2}(x)| = 0.11$. We then have

$$b_{i,j}|x - y| \leq |S_{i,j}(x) - S_{i,j}(y)| \leq c_{i,j}|x - y|$$

for $i, j \in \{1, 2\}$. Using again theorem 4 and theorem 5 we get $t \leq \dim_{\text{H}}(E) \leq s$ where t is the solution of $2(0.082)^t + 2(0.0625)^t = 1$ and s is the solution to $2(0.25)^s + 2(0.11)^s = 1$. Since $2(0.082)^{0.53} + 2(0.0625)^{0.53} = 0.991$ and $2(0.25)^{0.8} + 2(0.11)^{0.8} = 1.002$ we conclude that

$$0.53 < \dim_{\text{H}}(E) < 0.80.$$

Notes and references

The material presented in this section is based primarily on [Falconer, 1985], [Falconer, 1990] and [Hutchinson, 1981]. In the proof of theorem 3 we have mainly followed [Falconer, 1990] although we apply the mass distribution principle in a slightly more direct fashion. Example 5 is [Falconer, 1990, Exercise 9.9].

III CALCULATING ENTROPY AND DIMENSION OF LANGUAGES

In this section of the paper we explore the connection between entropy, crosswords and the Hausdorff dimension of certain self-similar sets.

Entropy of the crossword-language

In [Immink et al., 1998, Section VIII, A] the connection between the existence of crosswords and entropy is mentioned and the connection is explained by a calculation similar to the one we made in section I. However, the approach made in [Immink et al., 1998] is slightly different, in that the authors assume we have given $H^* = H/\log_2(|A|)$ where H is the binary entropy of the language that the crossword should contain. (The value H^* is known as the *relative entropy*.) Following Shannon we may assume a language of relative entropy H^* contains $|A|^{nH^*}$ sequences of length n , for large n . This may then be used to estimate, in the same manner as we did in section I, the number of two-dimensional crosswords to be

$$|A|^{n^2(2H^*-1)}. \quad (23)$$

This, in turn, may be rewritten to

$$\frac{(|A|^{nH^*})^{2n}}{|A|^{n^2}},$$

which is comparable to (1) from section I.

The authors of [Immink et al., 1998] further argues, that the sequences of words we find in the rows and columns of crosswords is not exactly natural language, because we are completely free to choose the words⁶ that form the sequences. Calculating, or, perhaps more accurately put, measuring the entropy of (semi-)natural languages is not a trivial task. Shannon mentions several possible approaches, including the measurement of n 'th order Markov probabilities and the entropy they induce as well as experiments where people try to read a text from which a number of letters have been removed. For the special (semi-)natural language that is found in crosswords, the authors of [Immink et al., 1998] suggests the following method for calculating the entropy. Let A be the alphabet and let $D = \{w_1, \dots, w_n\}$ be the dictionary of valid words, formed by taking any reasonable complete dictionary and appending \square to all words in it. We also want \square as well as $a\square$ for any $a \in A$ to be members of D . The valid sequences are now any sequence made as a concatenation of words from D . This may be seen as a constraint put on a channel, and we can calculate the Shannon capacity (see [Shannon, 1948, Part I, Section I], which is defined as

$$C = \lim_{n \rightarrow \infty} \frac{\log N(n)}{n},$$

where $N(n)$ is the number of valid sequences of length n . Grouping the words of D according to their length, we get a number of classes, $D_i = \{w \in d \mid |w| - 1 = i\}$ for $0 \leq i \leq M$ for some M . Note that D_i contains words whose length *exclusive* the additional \square is i and hence $D_0 = \{\square\}$. Writing $a_i = |D_i|$, that is a_i is the number of words of length i , we have

$$N(n) = N(n-1) + a_1N(n-2) + \dots + a_MN(n-M-1).$$

The solution to this difference equation will be dominated, for large n , by $c\gamma^n$ where c is a constant and γ is the largest real root of the polynomial

$$x^{-1} + a_1x^{-2} + \dots + a_Mx^{-M-1} - 1 = 0. \quad (24)$$

⁶A word, in this case, is a regular English word with an additional symbol \square appended to the end.

The calculation of the capacity C may now proceed, and we get

$$C = \lim_{n \rightarrow \infty} \frac{\log c \gamma^n}{n} = \lim_{n \rightarrow \infty} \frac{\log c + n \log \gamma}{n} = \log \gamma.$$

The entropy H^* of the crossword-language is thus estimated as $\log \gamma / \log |A| = \log_{|A|} \gamma$, where we have normalized with logarithm to the alphabet-size to get the relative entropy.

A connection to self similar sets

In section I and in the above subsection, we have worked with a crossword-language, that is, a language where valid sentences are concatenations of words from a dictionary. Although crosswords are made from finite sequences, we can consider the crossword-language to be a subset of A^∞ , namely as the set $L = \{w_1 w_2 \dots \mid w_i \in D\}$ of all concatenations of words from D . Valid rows and columns in the crossword are then simply prefixes of elements from L . By the standard method of representing numbers from the interval $[0, 1]$ in their $|A|$ -ary expansion, we may identify each sequence of L with a number in $[0, 1]$. In addition, each prefix x_1^n of a sequence $x \in L$ may be identified with a $|A|$ -ary interval in $[0, 1]$, that is, an interval of the form $[s/|A|^n, (s+1)/|A|^n]$. (See e.g. [Billingsley, 1965].) As an example of this interpretation, the Cantor set, C , may now be viewed as the set of all infinite sequences over an alphabet $A = 0, 1, 2$ made using only the letters 0 and 2. One IFS, whose invariant set equals the Cantor set, is $S = \{S_1, S_2\}$ with $S_1(x) = \frac{1}{3}x$ and $S_2(x) = \frac{1}{3}x + \frac{2}{3}$. In terms of the 3-ary expansion these similarities corresponds to shifting the fractional part to the left and inserting 0 respectively 2 on the first spot after the decimal point. Thus if $x = 0.2022 \dots \in C$ then $S_1(x) = 0.02022 \dots$ and $S_2(x) = 0.22022 \dots$.

Consider therefore the alphabet A and dictionary D as above. We may identify each of the letters of A with a integer, so that $A = \{0, \dots, |A| - 1\}$ is an alternative representation of the alphabet. Each finite sequence $x = x_1^n$ over A , can now be interpreted as a number expressed in base $|A|$, in particular $0.x$ is identified as the rational number $\sum_{j=1}^n x_j |A|^{-j}$. Recall also, that if $P = \{w_1, \dots, w_k\}$ is a prefix free set of finite sequences over A , then the set $\{[0.w\omega, 0.w\omega \mid w \in P]\}$, where ω is the highest numbered letter (i.e. $\omega = |A| - 1$), is a set of disjoint subintervals of $[0, 1]$. Let $k = |D|$, that is, $D = \{w_1, \dots, w_k\}$, and define $S_i : [0, 1] \rightarrow [0, 1]$ as

$$S_i(x) = \frac{1}{|A|^{|w_i|}} x + 0.w_i,$$

for any $i = 1, \dots, k$. It is clear that $S = \{S_1, \dots, S_k\}$ is a set of similarities, since $|S_i(x) - S_i(y)| = |A|^{-|w_i|} |x - y|$. Furthermore, the set D is a prefix free set, for D is made from a set of distinct sequences (regular words from a dictionary) to which we append a special letter, \square . It follows, that setting $S(A) = \cup_{i=1}^k S_i(A)$ for any set A , we find that

$$S([0, 1]) = \cup_{i=1}^k S_i([0, 1]) \tag{25}$$

is a disjoint union. Not surprisingly, the invariant set of the IFS $S = \{S_1, \dots, S_k\}$ is of interest. We argue that

$$E = \{0.\mathbf{w} \mid \mathbf{w} \in A^\infty \text{ and } \mathbf{w} = w_{i_1} w_{i_2} \dots \text{ with } w_{i_j} \in D\}$$

is the invariant set of S . We may say, that E is the set of numbers in $[0, 1]$ whose $|A|$ -ary expansion when interpreted as letters of A are allowable sequences from the crossword language over D . Assume $0.x \in E$. Then $S(\{0.x\}) = \cup_{i=1}^k S_i(0.x) = \cup_{i=1}^k \{0.w_i x\} \subseteq E$. Assume $0.x \in S(E)$.

Then $0.x = S_i(0.y)$ for some $0.y \in E$. Hence $0.x = 0.w_i y \in E$, and we conclude that E is the invariant set for the IFS.

It is now easy to calculate the Hausdorff and box-counting dimension of E . The equation (25) shows that the open set condition is satisfied with $]0, 1[$, and hence applying the dimension formula for self similar sets (section II, theorem 3) yields $\dim_H(E) = \dim_B(E) = s$ where s is the solution to

$$\sum_{i=1}^k \left(\frac{1}{|\mathcal{A}|^{w_i}} \right)^s = 1. \quad (26)$$

We can therefore conclude, that crossword-languages corresponds to certain subsets of $[0, 1]$, and that the Hausdorff dimension of these subsets can be calculated easily.

Hartley entropy and box counting dimension

The entropy concept that emerged in section I was defined on a subset L of A^* (see definition 2). However, by extending $c_L(n)$ to be also defined when $L \subseteq A^\infty$,

$$c_L(n) = |\{x \uparrow^n \mid x \in L\}|,$$

that is, the number of distinct prefixes of length n found in L , we can write

$$\tilde{H}(L) = \lim_{n \rightarrow \infty} \frac{\log_{|\mathcal{A}|} c_L(n)}{n}, \quad (27)$$

where $L \subseteq A^\infty$ is the crossword-language considered as a subset of all the infinite sequences. Since the crossword-language is closed under concatenation of finite sequences, we find that the definition of \tilde{H} given here is equivalent with the one given in section I. We now also recognize $\tilde{H}(L)$ to be the Hartley entropy. As it turned out, the value $\tilde{H}(L)$ was critical for the possibility of constructing crosswords. Recall from above that the crossword-language $L \subseteq A^\infty$ corresponds precisely to a set $E \subseteq [0, 1]$. The value of $c_L(n)$ is then equal to the number of intervals of the form $[s/|\mathcal{A}|^n, (s+1)/|\mathcal{A}|^n[$ which intersects with E . In other words, $c_L(n) = N_{\delta_n}(E)$ where $\delta_n = |\mathcal{A}|^{-n}$ (see definition 6). By the definition of the box-counting dimension and the fact that

$$\frac{\log c_L(n)}{n} = \frac{\log N_{\delta_n}(E)}{-\log_{|\mathcal{A}|} \delta_n}$$

we find $\tilde{H}(L)$ to be equal to $\dim_B(E)$. (All limits involved exists since $\dim_B(E)$ is guaranteed to exist by the dimension formula.)

These considerations lead us to the conclusion that the entropy of crossword-languages introduced in section I was in fact the Hartley entropy, and that it equals the box counting dimension of the corresponding subset of the unit interval (at least when both the entropy and the dimension exists). Furthermore, because of the self-similar nature of the subset of $[0, 1]$ the box counting dimension is known to exist and to be equal the Hausdorff dimension.

Connecting the dots

Assume γ is a solution to the equation (24), so that

$$\gamma^{-1} + a_1 \gamma^{-2} + \dots + a_M \gamma^{-M-1} - 1 = 0.$$

Recalling that a_i was the number of words of length i (excluding the \square symbol at the end), we find that for each i the expression $a_i \gamma^{-i-1}$ equals $\sum \gamma^{-|w|}$ where the sum is taken over all words w of D with $|w| = i + 1$. Combining for all values of i , the polynomial now take the form

$$\sum_{w \in D} \gamma^{-|w|} = 1,$$

or, by rearranging a bit,

$$\sum_{w \in D} \frac{1}{|A|^{(\log_{|A|} \gamma)^{|w|}}} = \sum_{i=1}^k \left(\frac{1}{|A|^{|w_i|}} \right)^{\log_{|A|} \gamma} = 1.$$

It is seen that when γ is the solution to the polynomial (24), then $\log_{|A|} \gamma$ is the solution to (26) as well.

The circle is now complete: we have shown that the calculation of the entropy of the crossword-language made in [Immink et al., 1998], corresponds to finding the solution of (26), that is, the similarity dimension. Since the crossword-language is self-similar, the similarity dimension equals the box counting dimension and the Hausdorff dimension.⁷

Empirical results

We now present a number of empirical results regarding the Hausdorff and box-counting dimension of crossword-languages. We have considered three dictionaries, two of them English and one Danish.

The calculation of the Hausdorff dimension has been made using the computer program shown in Appendix I. The method of calculation is the one outlined above and we have proceeded as follows. We have only considered words made from the 26 (resp. 29 letters) and ignored words or word-forms including numbers, apostrophes etc. Furthermore, all words have been converted to lower-case only. For each word w from the given dictionary we then form the special word $w\square$ and create a new dictionary like this: $D = \{\square, a\square, aback\square, abacus\square, \dots\}$. Based on the words of D we then calculate the similarity constants of the functions from the IFS, $c_w = |A|^{-|w|}$, and define the map $f(s) = \sum_{w \in D} c_w^s$. The equation $f(s) = 1$ is then solved by a simple numerical method. The results are shown below.

Dictionary	Words	Letters	Hausdorff dimension
English1	45.356	27	0.54
English2	134.477	27	0.61
Danish	360.331	30	0.64

Based on these (admittedly somewhat crude) measurements it seems that there should be no immediate shortage on crosswords. It would be interesting to compare these numbers to more detailed studies of the statistics and entropy of languages (both natural and semi-natural, as the crossword languages).

⁷We may add, that it is well known (e.g. from [Billingsley, 1965]) that Hausdorff dimension and the traditional, statistical entropy are closely related. In particular, the set of typical sequences for a stochastic process have Hausdorff dimension equal to the entropy of the process.

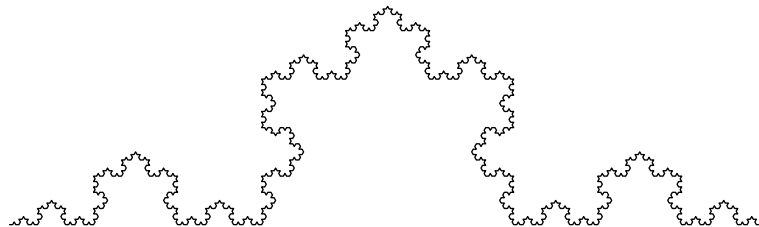
Notes and references

The subsection entitled “Entropy of the crossword-language” is based on [Immink et al., 1998]. In that article is also found an interesting passage relating to the remark on crossword made in Shannon’s 1948 article,

[E. Gilbert tells:] I didn’t understand that crossword example and tried to reconstruct his argument. That led to a kind of hand-waving “proof,” which I showed to Claude. Claude’s own argument turned out to have been something like mine. . . . Fortunately, I outlined my proof in the margin of my reprint of the paper (like Fermat and his copy of Diophantos). It went like this: . . .

The argument that followed matched the one given in [Immink et al., 1998], so we now have a quite good idea of what Shannon’s thoughts was on this subject.

The dictionaries used in the empirical results of the last subsection are: `English1`: The standard words file of the Debian Linux distribution; `English2`: The word list used in the spell checker `ispell`; and `Danish`: The word list “Den Store Danske Ordliste” collected and maintained by SSLUG (Skåne Sjælland Linux User Group).



The von Koch curve is an example of a two-dimensional self-similar set. It has Hausdorff and box counting dimension $\log 4 / \log 3$ and is also quite nice to look at.

REFERENCES

- Billingsley, P. *Ergodic theory and Information*. John Wiley & Sons, 1965.
- Falconer, K. *The geometry of fractal sets*. Cambridge University Press, 1985.
- Falconer, K. *Fractal Geometry - Mathematical Foundations and Applications*. John Wiley & Sons, 1990.
- Hutchinson, John E. Fractals and self similarity. *Indiana University Mathematics Journal*, 30(5): 713–747, 1981.
- Immink, K.A.S., P.H. Siegel and J.K. Wolf. Codes for digital recorders. *IEEE Trans. Inform. Theory*, 44(6):2260–2299, 1998.
- Ryabko, B. Y. Noiseless coding of combinatorial sources, Hausdorff dimension, and Kolmogorov complexity. *Problems of Inform. Trans.*, 22(3):170–179, 1986.
- Shannon, C.E. A mathematical theory of communication. Technical report, Bell System, 1948.

APPENDIX I

```

1  #!/usr/bin/perl
# ----- #
# language-dimension.pl - a program for estimating similarity dimension of
#                           IFS's induced by dictionaries
# ----- #

use strict;
use warnings;
$|=1;

10  my $precision = 0.001; # Stop when within $precision of the solution
    my $s         = 0.5;   # First dimension to try
    my @s         = (0,2); # Interval to search

# ----- #

    my %letters;
    my %words;

20  while(<>) {
        chomp;
        next unless /^[a-zA-Z]*$/;
        $_ =~ tr/A-Z/a-z/;
        s/$/ /;
        $letters{$_}++ foreach split//,$_;
        $words{$_}++;
    }

# ----- #

30  my $letters = (keys %letters)+0;
    my $words  = (keys %words )+0;
    print "Letters: $letters (".(join"",(keys%letters)).")\n";
    print "Words  : $words\n";

    sub sum {
        my $sum = 0;
        foreach my $word (keys %words) {
            $sum += (1/$letters)**(length($word)*$s);
40         }
        return $sum;
    }

    print "Trying dimensions: ";
    while (1) {
        printf "%6.4f ", $s;
        my $sum = sum();
        last if abs($sum-1) < $precision;
        if ($sum<1) { $s[1]=$s } else { $s[0]=$s }
50         $s = $s[0]+($s[1]-$s[0])/2;
    }

    print "\n";
    print "Estimated dimension: $s\n";

# ----- #

```