Lecture Notes in Mathematics

Edited by A. Dold and B. Eckmann

1135

Number Theory

New York 1983–84

Edited by D. V. Chudnovsky, G. V. Chudnovsky, H. Cohn and M. B. Nathanson

Springer-Verlag
Berlin Heidelberg New York Tokyo

Springer Lecture Notes in Mathematics #1135
New York Number Theory Seminar 1983-84

HAUSDORFF DIMENSION OF SETS ARISING IN NUMBER THEORY

Richard T. Bumby
Rutgers University

Introduction. Hausdorff gave his name to a process for constructing d-dimensional measure for arbitrary real numbers d in 1919 [10]. Some of the developments since that time can be found in the books of Billingsley [3], Mandelbrot [11], and Rogers [12]. I do not want to get involved in the technicalities of the measure theory here; so I will simply encourage you to think of d-dimensionality as meaning that if a set is scaled by a factor of k, then the measure is multiplied by k^d. The d-th power of the diameter has this property, so it is natural to compare d-dimensional measures with the set function $X \to (\text{diam } X)^d$. Similarly, you should think of sets as subsets of the real line as the examples I wish to present are sets of real numbers.

The notion of "measure zero" is easier than the general notion of measure. A set has measure zero if it may be covered by a union of sets which is arbitrarily small in the appropriate sense. In particular, a set X will be said to have measure zero in dimension d if for all ε, there is a sequence of sets U_i with

$$X \subseteq \bigcup U_i \qquad \Sigma(\text{diam } U_i)^d < \varepsilon$$

Notice that as $\varepsilon \to 0$ so does $\max(\text{diam } U_i)$. The notion of "finite measure" which is then suggested is that there is a number M such that for all $\delta > 0$ there is a sequence U_i with

$$X \subseteq \bigcup U_i \qquad \text{diam } U_i \leq \delta \quad \Sigma(\text{diam } U_i)^d \leq M.$$

If X does not have finite measure in a given dimension it will be said to have infinite measure in that dimension. The Hausdorff dimension is characterized by

Proposition 1. If X has finite measure in dimension d, then it has measure zero in all larger dimensions. If X has non zero measure in dimension d, then it has infinite measure in all smaller dimensions. Thus each set X has a dimension $\dim(X)$ such that X has measure zero in dimension $d > \dim(X)$ and infinite measure in dimension $d < \dim(X)$.

Proof. Exercise based on the observation that r^d is a decreasing function of d if $0 < r < 1$.

Note that the emphasis on covers means that one is essentially dealing with an outer measure. This has led to some apparent pathologies. The examples given here are actually rather tame; in particular, they have finite, non-zero measure in their Hausdorff dimension.

There are two common types of representation of numbers. The first is the decimal or, more generally, the base g representation; and the second is the continued fraction. Both have been studied in the context of Hausdorff dimension, and I will give an example of the determination of the Hausdorff dimension of a set defined by the properties of each of these representations.

The CLASSICAL CANTOR SET, here called CCS, consists of those x, $0 \leq x \leq 1$ having a base 3 expansion without the digit 1. Since

$$CCS = c_0(CCS) \cup c_2(CCS) \quad \text{where} \quad c_i(x) = \frac{x+i}{3} \tag{1}$$

a d-dimensional measure would assign each $c_i(CCS)$ measure 3^{-d} times the measure of CCS. Since each of these is half of CCS, it seems that $d = \log 2/\log 3$ is the only reasonable choice for the dimension. This was proved by Hausdorff [10], but the generalization to sets defined by missing digits in the base g expansion came much later (see Best [2] and Volkmann [14]). I will give a proof in the next section.

The set E_2 consists of all numbers having infinite continued fraction expansions with partial quotients 1 or 2. In particular, every element x of E_2 satisfies

$$\frac{1+\sqrt{3}}{2} \leq x \leq 1 + \sqrt{3}$$

and the part of E_2 with zeroth partial quotient equal to i, $i = 1$ or 2, is selected by the map

$$e_i(x) = i + 1/x.$$

Since e_i is not an affine map, one cannot use the equation

$$E_2 = e_1(E_2) \cup e_2(E_2)$$

in the same way as one used (1) to guess the dimension.

Techniques for bounding the dimension of E_2 have been known for some time. In 1941, I.J. Good proved

$$.5194 \leq \dim E_2 \leq .5433 \tag{2}$$

More recently, I [5] proved a general theorem characterizing the dimensions of such sets. The method was sufficiently constructive that it was possible to compute upper and lower bounds on the dimension of E_2. Unfortunately, the method was so slowly convergent that the results did not strengthen (2). Meanwhile, Cusick [6] found that $\dim E_2$ arose naturally in another problem. This caused a renewed interest in finding a computation of $\dim E_2$ that was simultaneously naive, so that it could be easily explained, and accurate, so that it would be useful. The method sketched here seems to satisfy both requirements. The idea of monotocity which was so valuable in Good's calculation has been incorporated into my algorithm. It required only a few minutes on a home computer to prove

$$.5312 \leq \dim E_2 \leq .5314. \tag{3}$$

The proof of (3) will be sketched in the third section of this paper.

There is another example to which these methods apply which should be mentioned although I shall give no details. Start with three circular discs of radii r_1, r_2, r_3 in the plane each two of which are externally tangent. The complement of these discs has two components. The closure of the bounded component is a "triangle" whose sides are circular arcs. These triangles are parameterized by the numbers r_1, r_2, r_3. A classical construction determines a unique circular disc inscribed in this triangle. Removal of this disc leaves three smaller triangles in place of the given triangle. Iteration of this construction gives a Cantor set construction. The underlying geometry is now inversive geometry in the plane, or one-dimensional complex projective geometry. The Hausdorff dimension of this set has been determined by D.W. Boyd [4].

The arithmetic significance of this has been noted by A.L. Schmidt [13]. A complex continued fraction can be described by this construction, and a dual construction for subdividing the disc (marked with its three points of tangency) which was removed from the triangle.

<u>The dimension of CCS</u>. The cover of CCS by the intervals of length 3^{-n} occurring in its construction has $\Sigma(\text{diam } U_i)^d = 1$ for all n if $d = \log 2/\log 3$. It follows that CCS has finite measure in this dimension, hence measure zero in all larger dimensions. If $C \subset S$ can be shown not to have measure zero in dimension $d = \log 2/\log 3$ it will follow that CCS has dimension d. Part of that result is contained in

Proposition. If each U_i is of the form $[a/3^n, a+1/3^n]$ where a has only digits 0 and 2 base 3, then

$$CCS \subseteq \cup U_i \Longrightarrow \Sigma(\text{diam } U_i)^d \geq 1.$$

Remarks. The value of n is allowed to depend on U_i, and there may be infinitely many such intervals, so this is different from the observation used to show that CCS has finite measure in dimension d. Conventional wisdom suggests that the U_i could be replaced by open sets with a small effect on the measure and then compactness could somehow be used to get the desired conclusion or something just as good. Indeed, the classical method for finding dimensions of subsets of the real line used a theorem of Gillis [7] whose proof involved the Heine-Borel theorem. I will prove a "little man who isn't there" theorem which is the contrapositive of the given statement by an argument which resembles a proof of the Heine-Borel theorem attributed to Besicovitch (see Hardy [9], section 106).

Proof. Start with a collection of intervals U_i of the required type with $\Sigma(\text{diam } U_i)^d < 1$. For the purpose of this proof, such a collection will be called "deficient".

Clearly no deficient collection can contain [0,1]. Hence the intervals can meet only one of c_0(CCS), c_2(CCS). The sum then splits into two pieces, so at least one piece is less than 1/2. Select c_i so that the sum associated to c_i(CCS) is less than 1/2. Then the image of these intervals under c_i^{-1} is a deficient collection.

This sequence of choices determine a nest of intervals in the construction of CCS. The intersection of these intervals is our "little man". The above constructions shows that the intervals $[a/3^n, a+1/3^n]$ which contain him can not be in our collection of intervals U_i. On the other hand, these are the only intervals which we have allowed as U_i which can contain him. So, he is not there!

To show that these special intervals suffice to determine the Hausdorff dimension, associate an arbitrary set U_i with the smallest interval $[a/3^n,(a+1)/3^n]$ containing $U_i \cap$ CCS. It is easily seen that this interval is no more than three times the length of U_i.

A more delicate analysis will show that our proposition holds for arbitrary U_i. This analysis has been employed by Wegmann [15] to determine the measure in dimension $\log(g-1)/\log g$ of the set of numbers missing the digit a in their base g expansions.

The dimension of E_2. The mappings e_1 and e_2 relating E_2 to its two parts are not linear. They cannot be used to give the lengths of the construction intervals as easily as was done for CCS. However, they are fractional linear functions so they preserve the projective geometry of the line. Each ratio AB/AC can be written as a cross ratio $\frac{AB \cdot PC}{AC \cdot PB}$ where P is the point at infinity. To determine the effect when this ratio is mapped by e_i^{-1}, it suffices to locate $e_i^{-1}(P)$ and compute the appropriate cross ratio. Over several such steps, the point at infinite goes first to 0, then to a value between -1 and $-1/3$ and eventually to a neighborhoood of

$$\{x \mid -\tfrac{1}{x} \in E_2\}.$$

The lengths of various subintervals of one of the construction intervals as a ratio of the length of that interval can be expressed in terms of lengths of the corresponding subintervals of the "template" $I = [\frac{\sqrt{3}+1}{2}, \sqrt{3}+1]$ by using the inverse image of the point at infinity under the mapping sending the template onto the given construction interval. This function is a continuous function of the point corresponding to infinity. If, as in [5], this value is written as $-1/\lambda$, the values of λ in a neighborhood of E_2 determine all other values.

If E_2 has a d-dimensional measure, then the ratio of any projective transform of E_2 to the d-th power of the length of that transform of I will be a continuous function of λ where $-1/\lambda$ is the preimage of infinity for that mapping. Use of $e_1(E_2)$ and $e_2(E_2)$ to compute the measure gives rise to a subdivision operator

$$(Tf)(\lambda) = \left[(2-\sqrt{3})\ \left(\frac{1+\lambda(\sqrt{3}+1)}{1+\lambda\sqrt{3}}\right)\right]^d f(1+\frac{1}{\lambda}) + \left[(2-\sqrt{3})\ \left(\frac{2+\lambda(\sqrt{3}+1)}{2+\lambda(\sqrt{3}+3)}\right)\right]^d f(2+\frac{1}{\lambda}).$$

[This formula suffered a misprint (the insertion of a spurious '+') on p. 201 of [5] which is corrected here.] Any measure would then be a positive invariant function, i.e. a positive eigenfunction with eigenvalue 1. The result of [5] is that the classical Perron-Frobenius theory of non-negative matrices holds for T. The eigenfunction can then be used to show that the measure of E_2 is zero if the eigenvalue is less than 1 and infinite if the eigenvalue is greater than 1. The eigenvalue is a continuous decreasing function of d so there will be a dimension in which the eigenvalue is equal to 1, and in this dimension the eigenfunction gives the ratio of the measure of each part of E_2 to the d-th power of the length of the corresponding construction interval. This eigenfunction is unique up to a scale factor. The usual properties of Haar measure on groups [16] seem to have analogies here.

How can this be used to compute the dimension as accurately as I have claimed? The first ingredient is a corollary of the Perron-Frobenius theory.

Proposition. A necessary and sufficient condition that the spectral radius of T be greater than or equal to 1 is that there exist a positive function f with $Tf \geq f$. Dually, the spectral radius of T is less than or equal to 1 if and only if there is a positive function with $Tf \leq f$.

Proof (sketch). The eigenfunction establishes the necessity. To prove sufficiency, note that $\{f: Tf \geq f \geq 0\}$ is taken into itself by the positive operator T. Combining this with the proof of existence of an eigenfunction shows that the eigenfunction has this property.

The second ingredient is the observation that T is conjugate to the operator C defined by

$$(Cf)(\lambda) = \lambda^{-2d} f(1+\frac{1}{\lambda}) + \lambda^{-2d} f(2+\frac{1}{\lambda}).$$

This could be based on the fact that the continuant to the power (-2) is an estimate of the length of the a construction interval for a Cantor set defined by the continued fraction expansion. (See Cusick [6] for properties of continuants.) Alternatively one could give an explicit multiplication operator M such that MC = TM (exercise).

The functions $g_\alpha(\lambda) = (\alpha\lambda+1)^{-2d}$ have special properties for the operator C. Exercise: compute $Cg_\alpha(\lambda)$. I won't spoil the surprise by giving the answer. It will follow that the eigenfunction has the form

$$\int g_{\alpha}(\lambda)\ d\mu(\alpha) \tag{4}$$

(whatever this means!). We shall only require the following corollary which could be obtained directly from the definition of C.

Proposition. There is a cone of positive decreasing functions stable under C. In particular, there is a constant $b < 0$ such that

$$\{f:\ 0 \geq f'/f \geq b\}$$

is stable under C.

Proof: Exercise.

Remark. There is much more to (4) than will be needed here. It seems to establish some sort of self-duality of C. In particular the measure μ appears to be the d-dimensional Hausdorff measure on E_2 that we have been seeking. For a related result on a different operator related to the continued fraction expansion see Babenko [1].

A method of producing a function satisfying $Cf \geq f$ in this case is to take

$$f_0 = 1,\ f_{n+1} = f_0 \vee Cf_n \quad (n \geq 0)$$

[Here $(f \vee g)$ stands for the function with $(f \vee g)(\lambda) = \max(f(\lambda), g(\lambda))$.]

Proposition. $f_{n+1} \geq f_n$. Hence the following are equivalent:

(i) $Cf_n \geq f_n$

(ii) $Cf_n \geq f_0$

(iii) $(Cf_n)(1+\sqrt{3}) \geq 1$.

Proof: $f_1 = f_0 \vee Cf_0 \geq f_0$. Now if $f_n \geq f_{n-1}$, then $f_{n+1} = f_0 \vee Cf_n \geq f_0 \vee Cf_{n-1} = f_n$. Induction on this shows that $f_n \geq f_0$ so that (i) $\Longrightarrow$ (ii). Conversely (ii) $\Longrightarrow$ $Cf_n = f_{n+1} \geq f_n$ which is (i). Since Cf_n is a decreasing function, it exceeds 1 everywhere on I if it exceeds 1 at the right endpoint. Thus (ii) and (iii) are equivalent. (I must apologize for stating the proposition as if it were specific to this one example. The proof is easy enough that other applications should afford no difficulty.)

It is easy enough to do this with $n = 10$. It takes just a couple of minutes to compute $(Cf_n)(1+\sqrt{3})$ recursively on a home computer. (I also have a version of this algorithm for a programmable calculator - but a calculation with $n = 10$ takes several days on such a machine.) Much larger values of n would give only slight improvements in the result, but would take much longer as the number of steps is proportional to 2^n.

In actuality, (3) required a further improvement on this method. The details can be skipped here as better methods are likely to be found. I should leave you with some hints about the methods I used since they lead to some results about the functions g_α which are interesting in their own right.

Exercises: (1) Determine conditions for g_α/g_β to be increasing.

(2) Let $h_\alpha = g_{1+\frac{1}{\alpha}} + g_{2+\frac{1}{\alpha}}$. Express Ch_α in terms of the h_β.

(3) Determine conditions for h_α/h_β to be increasing.

It will follow from the exercises that the operator which multiplies by h_β, applies C, and then divides by h_β will have a stable cone of monotonic functions if $\beta = 1+\sqrt{3}$ or $\beta = (1+\sqrt{3})/2$. These are the operators used to prove (3).

Conclusions. The sets considered here have their roots in arithmetic, but the theoretical tools introduced to compute their Hausdorff dimensions should have broader interest and application. In particular, the relation of the Hausdorff dimension to the spectral radius of the subdivision operator provides a means of eliminating ad hoc estimates, thereby sharpening the calculations. The use of monotonicity to allow inequalities of functions to be tested by finite numerical calculations does not seem to have a place in the numerical analysis arsenal. It bears further study. The "spectral analysis" given by equation (4) illustrate a self-duality which seems to be present also for the circle-packing example. This is likely to be an important structure.

References

1. K.I. Babenko. On a problem of Gauss; Dokl. Akad. Nauk SSSR 238 (1978), 1021-1024 Soviet Math. Dokl. 19 (1978), 136-140.

2. E. Best. On sets of fractional dimension, III; Proc. London Math. Soc. (2) 47 (1942), 436-454.

3. P. Billingsley. Ergodic theory and information. Wiley, New York 1965.

4. D.W. Boyd. The residual set dimension of the Apollonian packing; Mathematika 20 (1973), 170-174.

5. R.T. Bumby. Hausdorff dimensions of Cantor sets; J. reine angew. Math. 331 (1982), 192-206.

6. T.W. Cusick. Continuants with bounded digits, I; Mathematika 24 (1977), 166-127. __________, II; ibid. 25 (1978), 107-108.

7. J. Gillis. Note on a theorem of Myrberg; Proc. Camb. Phil. Soc. 33 (1937), 419-424.

8. I.J. Good. The fractional dimension theory of continued fractions; Proc. Camb. Phil. Soc. 37 (1941), 199-228.

9. G.H. Hardy. A course of pure mathematics (10^{th} edition). Cambridge University Press, Cambridge 1952.

10. F. Hausdorff. Dimension und äusseres Mass; Math. Ann. 79 (1919), 157-179.

11. B. Mandelbrot. Fractals: form, chance, and dimension. Freeman, San Francisco 1977.

12. C.A. Rogers. Hausdorff measures. Cambridge University Press, Cambridge 1970.

13. A.L. Schmidt. Ergodic theory for complex continued fractions; Monatsh. Math. 93 (1982), 39-62.

14. B. Volkmann. Über Hausdorffsche Dimensionen von Mengen die durch Ziffereigenschaften charakterisiert sind I; Math. Z. 58 (1953), 284-287. _______, II; ibid. 59 (1953), 247-254. _______, III; ibid. 59 (1953), 259-270. _______, IV; ibid. 59 (1954), 425-433. _______, V; ibid. 65 (1956), 389-413.

15. H. Wegmann. Das Hausdorff-Mass von Cantormengen; Math. Ann. 193 (1971), 7-20.

16. A. Weil. L'intégration dans les groupes topologique et ses applications. Hermann, Paris 1951.