THE CANTOR SET AND A GEOMETRIC CONSTRUCTION

L'Enseignement Mathématique, t. 35 (1989), p. 41-49

THE CANTOR SET AND A GEOMETRIC CONSTRUCTION

by Marco PAVONE

INTRODUCTION

The Cantor ternary set consists of all those real numbers x in $[0, 1]$ which have a ternary expansion $x = \sum_{n=1}^{\infty} a_n/3^n$ for which a_n is never 1. Equivalently, C can be obtained in a purely geometrical fashion by first removing from $[0, 1]$ the middle third $(1/3, 2/3)$, then removing the middle thirds $(1/9, 2/9)$ and $(7/9, 8/9)$ of the remaining intervals, and so on (C will be exactly the complement of the countable union of the removed intervals). If $x = \sum_{n=1}^{\infty} a_n/3^n$ is in C, the geometric interpretation of its ternary expansion is that x is the unique point in $[0, 1]$ which is reached by first staying to the left or to the right of $(1/3, 2/3)$ if $a_1 = 0$ or $a_1 = 2$ respectively, then staying to the left or to the right of the next removed interval if $a_2 = 0$ or $a_2 = 2$ respectively, and so on. It follows from the construction that C is a nowhere dense closed subset of $[0, 1]$.

A well known property of C is that any real number in $[0, 2]$ can be written as the sum of two numbers in C. The purpose of this note is to give an elementary proof of $C + C = [0, 2]$ which only uses the geometric definition of C. A refinement of the proof shows in fact that for any k in $[0, 2]$ there exists either a finite or an uncountable number of pairs x, y from C such that $x + y = k$. We also discuss the analogy between this decomposition result and certain properties of continued fractions.

THE GEOMETRIC CONSTRUCTION

We set, as usual, $C \times C = \{(x, y) \in \mathbf{R}^2 : x, y \in C\}$. Then $C + C = [0, 2]$ can be geometrically restated as

(*) for any k in $[0, 2]$ the line $x + y = k$ intersects $C \times C$ in at least one point.

Let's agree to call a line segment in $\mathbf{R}^2$ "horizontal" or "vertical" if it is parallel or perpendicular to the line $y = x$ respectively. Consider a sequence $L_0, L_1, L_2, \ldots$ of continuous polygonal curves in $\mathbf{R}^2$ with the following properties (see fig. 1-3):

(a) L_n is contained in $[0, 1] \times [0, 1]$ for all n, and is composed by horizontal and vertical segments only.

(b) The vertices of L_n belong to $C \times C$ for all n.

(c) The endpoints of L_n are $(0, 0)$ and $(1, 1)$ for all n.

(d) Each L_n contains 3^n horizontal segments, each of which has length $2^{1/2}/3^n$.

(e) For all n, and for any k in $\{0, 2/3^n, 4/3^n, \ldots, 2\}$ the line $x + y = k$ contains a vertical segment of L_n.

(f) For all n, and for any k not in $\{0, 2/3^n, 4/3^n, \ldots, 2\}$ the line $x + y = k$ meets at most one horizontal segment of L_n.

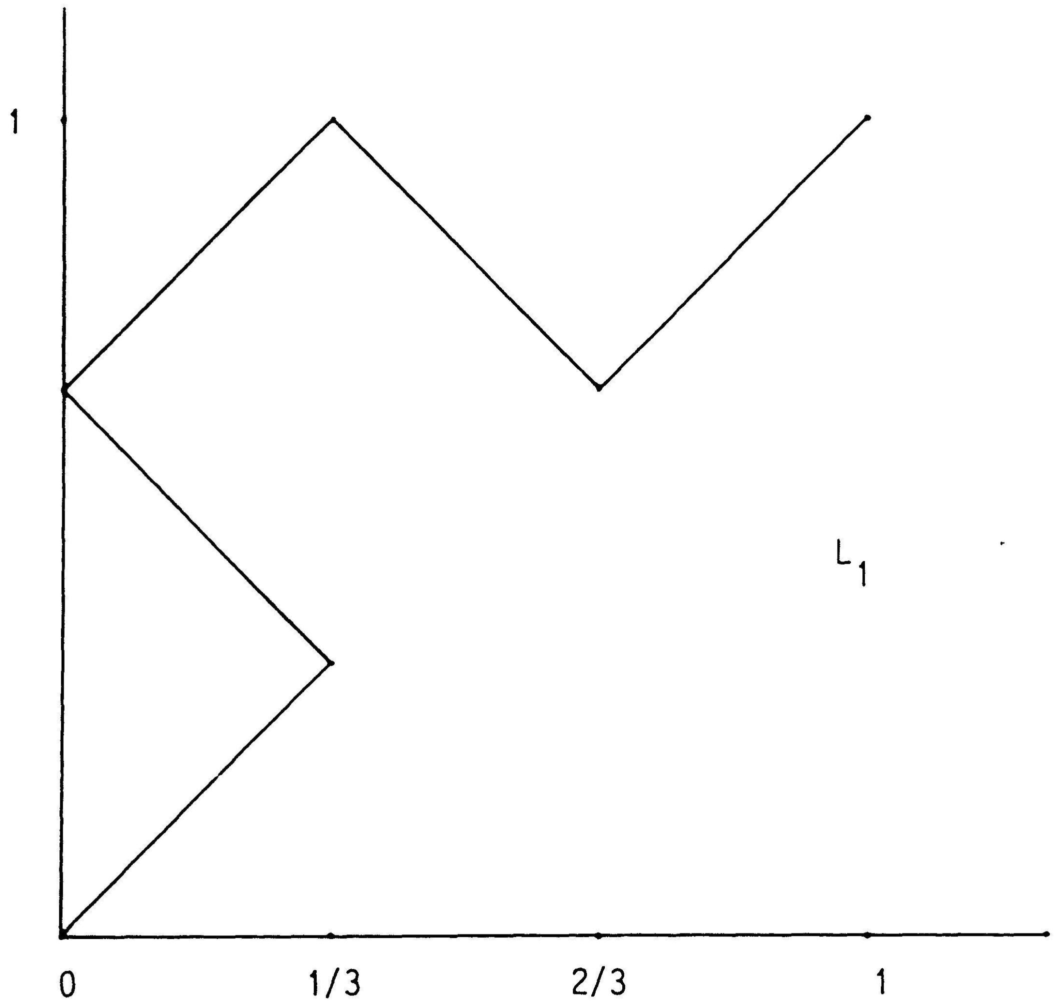

FIGURE 1

Suppose first that such a sequence exists. Then property (*) is satisfied. Indeed, fix k in $[0, 2]$ and let r denote the line $x + y = k$. If k is in $\{0, 2/3^n, 4/3^n, ..., 2\}$ for some n, then r meets $C \times C$ by (e) and (b); otherwise, for any positive integer n there exists by (f) a unique horizontal segment of L_n that meets r. This implies, by (d) and (b), that $\operatorname{dist}(r, C \times C) < 2^{1/2}/3^n$ for all positive integers n, that is, $\operatorname{dist}(r, C \times C) = 0$. Then r meets $C \times C$ by a standard compactness argument (I recall that C is a closed subset of $[0, 1]$).

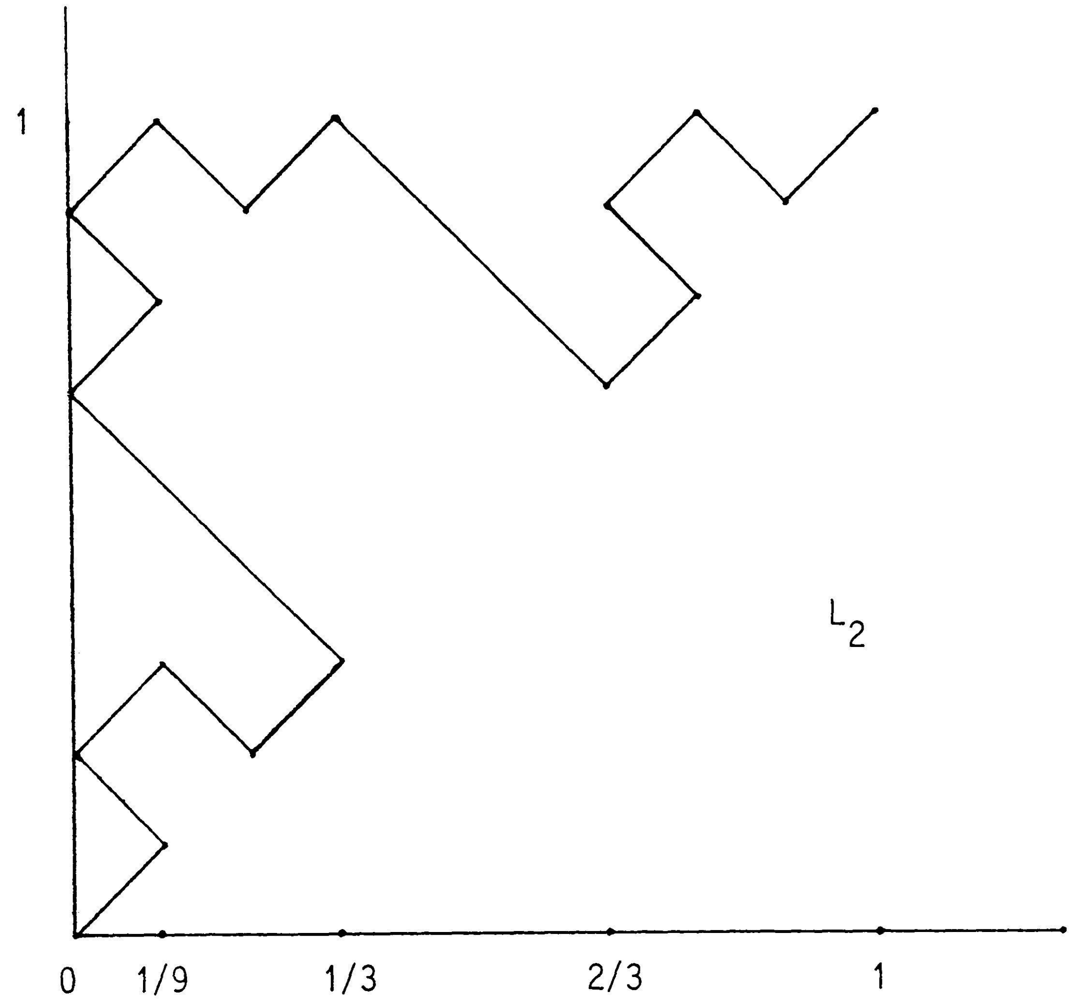

FIGURE 2

We now proceed to the heart of the argument, that is the construction of the sequence $\{L_n\}_n$. All we need is in fact the first step of an induction process. Let L_0 be the line segment with endpoints (0, 0) and (1, 1), and let L_1 be the polygonal with vertices (0, 0), (1/3, 1/3), (0, 2/3), (1/3, 1), (2/3, 2/3) and (1, 1) (see fig. 1). In general, let L_{n+1} be the curve obtained from L_n by performing on each horizontal segment of L_n the same modification that was performed on L_0 to get L_1. In other words, we replace the generic horizontal segment of L_n with endpoints (x, y) and $(x+1/3^n, y+1/3^n)$ by the polygonal passing through the points

$$(x, y), \quad (x+1/3^{n+1}, y+1/3^{n+1}), \quad (x, y+2/3^{n+1}), \quad (x+1/3^{n+1}, y+1/3^n), \\ (x+2/3^{n+1}, y+2/3^{n+1}) \quad \text{and} \quad (x+1/3^n, y+1/3^n)$$

(see fig. 2 and 3). It is then apparent that $\{L_n\}_n$ satisfies the hypotheses (a), ..., (f) stated above.

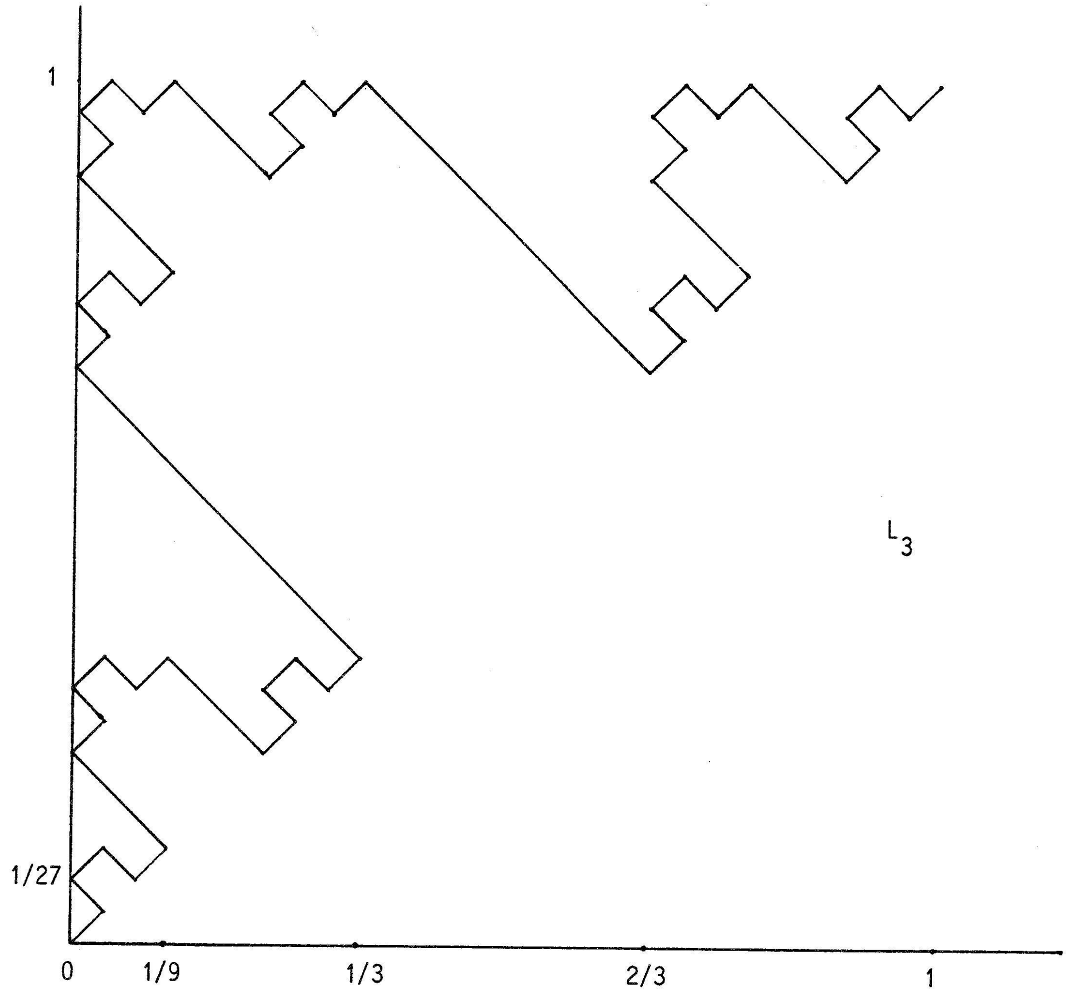

FIGURE 3

An easy modification of the previous construction gives us more information on the way a number in $[0, 2]$ can be written as the sum of two numbers in C. For every map μ from $\mathbf{N}\setminus\{0\}$ into $\{0, 2\}$ we construct a sequence $\{L_n^{(\mu)}\}_n$ of polygonal curves with properties (a), ..., (f). The idea is simply to add to the previous construction a choice between "left" and "right" at every step of the induction. What one ends up with is exactly a two-dimensional version of the geometric construction of the Cantor ternary set. We proceed as follows.

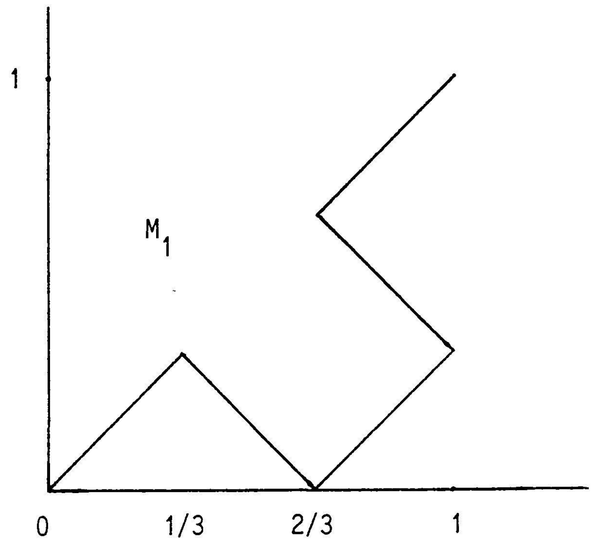

FIGURE 4

Let M_1 be the mirror image of the curve L_1 with respect to the line $y = x$ (see figure 4). If μ is a map from $\mathbf{N}\backslash\{0\}$ into $\{0, 2\}$, we define $L_0^{(\mu)} = L_0$, and for any nonnegative integer n we let $L_{n+1}^{(\mu)}$ be the polygonal obtained from $L_n^{(\mu)}$ by replacing each horizontal segment of L_n by a (normalized) copy of L_1 or M_1, according to whether $\mu(n+1) = 0$ or $\mu(n+1) = 2$ respectively. For example, if $\mu = \{0, 0, 0, ...\}$, we obtain our original sequence $\{L_n\}_n$ (fig. 1-3), and for $\mu = \{2, 2, 2, ...\}$ we get its mirror image with respect to the line $y = x$. For $\mu = \{0, 2, 0, 2, ...\}$, we obtain castle-like polygonals as in figure 5.

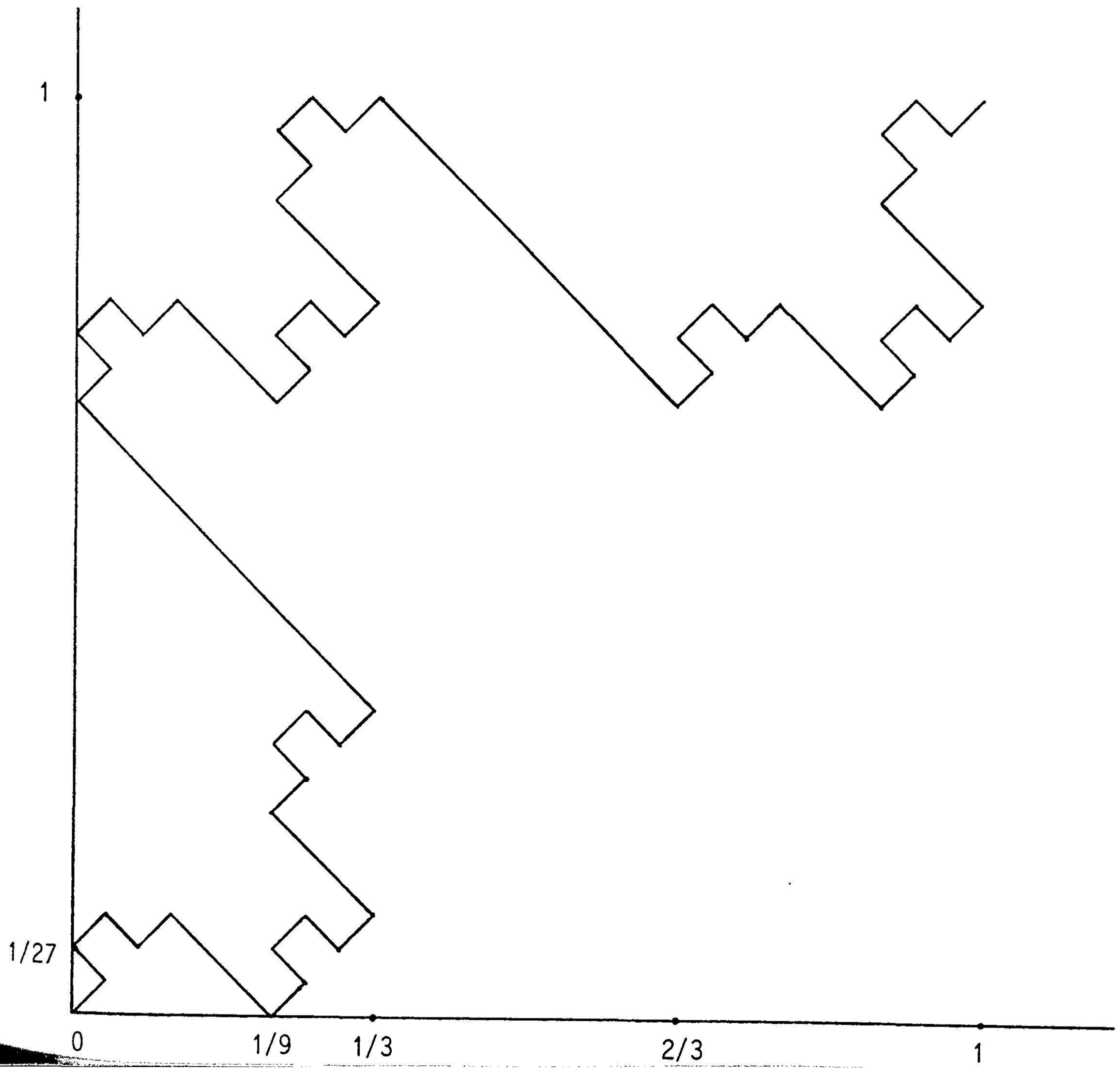

FIGURE 5

For all μ let $L^{(\mu)}$ denote the uniform limit of the curves $L_n^{(\mu)}$, $n = 0, 1, \ldots$. Then $L^{(\mu)}$ is a continuous curve in $[0, 1] \times [0, 1]$ with endpoints $(0, 0)$ and $(1, 1)$, and with the property that, for any k in $[0, 2]$, the line $x + y = k$ intersects $L^{(\mu)}$ in some point of $C \times C$. Viceversa, given any point (x, y) in $C \times C$, there is some sequence μ such that (x, y) lies on $L^{(\mu)}$.

To see this, note that the ternary subdivision of $[0, 1]$ that generates C produces a corresponding subdivision of $[0, 1] \times [0, 1]$ that generates $C \times C$. At the n-th step, the subset G_n of $[0, 1] \times [0, 1]$ that contains points of $C \times C$ is the union of 4^n squares (the black squares in figure 6 for $n = 3$). It is clear that G_n contains the vertices of the curves $L_n^{(\mu)}$ for all μ (compare figures 3 and 6). The conclusion is now immediate.

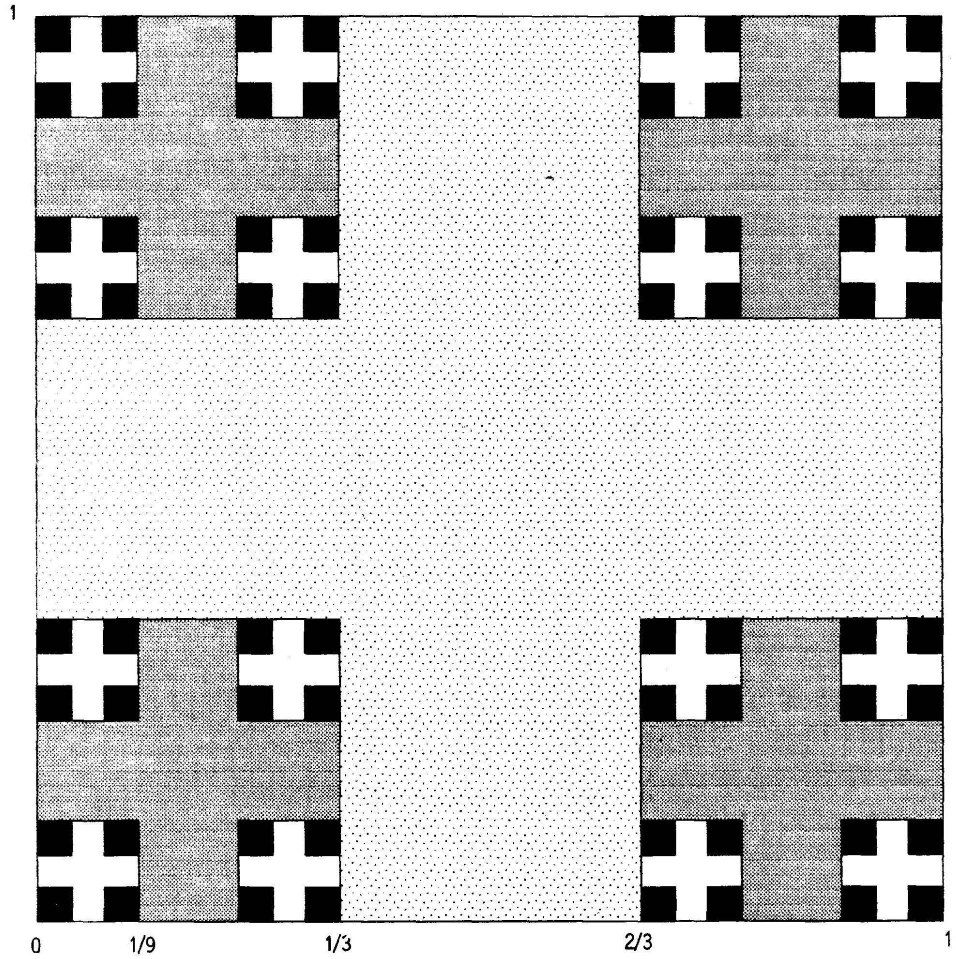

FIGURE 6

Note that if $\mu^\wedge$ is the sequence obtained from μ by turning all the 0's in 2's and viceversa, then the line $x + y = k$ intersects $L^{(\mu)}$ in a point (x, y) if and only if it intersects $L^{(\mu^\wedge)}$ in a point (y, x); in other words, $\mu^\wedge$ does not give us any new information on the decomposition of k as a sum of numbers in C. We shall therefore restrict our attention to sequences μ with $\mu(1) = 0$ (i.e. to curves $L^{(\mu)}$ above the line $y = x$).

Fix $k = 2h$ in $[0, 2]$, $h > 0$, and let $h = \sum_{n=1}^{\infty} a_n/3^n$ be the unique infinite ternary expansion of h. We claim that the equation $x + y = k$ has a finite or an uncountable number $S(k)$ of solutions in $C \times C$ according to whether the cardinality $c(k)$ of the set $\{n \in \mathbf{N}\backslash\{0\}; a_n = 1\}$ is finite or infinite respectively. In fact, the exact formula is $S(k) = 1$ if $c(k) = 0$ or 1, and $S(k) = 3(2^{c(k)-2})$ otherwise.

Let r be the line $x + y = k$, and let n be any positive integer. With the notation set above, and with the help of figure 6, it is easy to see that $a_n = 1$ if and only if G_n meets r in twice as many squares than G_{n-1}. Equivalently, $a_n = 1$ if and only if, for all μ, r meets $L_{n-1}^{(\mu)}$ in the middle third of one of its horizontal segments; in other words, $a_n = 1$ if and only if at the n-th step of the construction the curves $L_n^{(\mu)}$ meet r in twice as many points than the curves $L_{n-1}^{(\mu)}$. If $a_n \neq 1$, the choice between $\mu(n) = 0$ and $\mu(n) = 2$ at the n-th step does not produce any new intersection point. This shows that $c(k)$ is finite or infinite depending on whether r meets the curves $L^{(\mu)}$ in a finite or an uncountable number of points, and our claim is proved.

Example. If $k = 2h = 28/27$ ($h = 0.11122...$ in ternary form, with 2 repeated infinitely often), then $S(k) = 6$ and the possible decompositions are (in ternary form) $k = 1 + 0.001$, $k = 0.222 + 0.002$, $k = 0.221 + 0.01$, $k = 0.21 + 0.021$, $k = 0.202 + 0.022$ and $k = 0.201 + 0.1$.

In the case where $c(k)$ is infinite, we saw that each new occurence of 1 in the sequence $\{a_n\}_n$ produces a new choice between $\mu(n) = 0$ and $\mu(n) = 2$. In terms of the decomposition $k = x + y$, with $x = \sum_{n=1}^{\infty} b_n/3^n$ and $y = \sum_{n=1}^{\infty} c_n/3^n$, this corresponds precisely to choosing $b_n = c_n = 0$ if $a_n = 0$, $b_n = c_n = 2$ if $a_n = 2$, and finally $b_n = 0$ and $c_n = 2$ ($b_n = 2$ and $c_n = 0$) if $a_n = 1$ and $\mu(n) = 0$ ($\mu(n) = 2$). An interesting case is $k = 1$, that is, $h = 0.1111....$ In this case, if $1 = x + y$ is the decomposition determined by the choice of some sequence μ, then one has precisely $x = \sum_{n=1}^{\infty} \mu(n)/3^n$.

Remark. The construction of the sequence $\{L_n\}_n$ (fig. 1-3) is similar to the ones which define by induction the continuous nowhere-differentiable function on $[0, 1]$ or an infinite homogeneous tree with finite degree. They all provide examples of those geometric objects which are nowadays called fractals. A fractal has the property that each of its portions looks exactly like a reduced copy of the whole thing. This "homogeneousness" property has often an algebraic counterpart: in the case of the Cantor ternary set, the N-th step of its geometric construction corresponds to the fact that every number of the form $\sum_1^{N+1} a_n/3^n$, $a_n \in \{0, 1, 2\}$ is obtained from the number $\sum_1^N a_n/3^n$ by making a choice between $a_{n+1} = 0$, $a_{n+1} = 1$ and $a_{n+1} = 2$. The crucial point is that the nature of this choice does not depend on the number and does not depend on N. In $\mathbf{F}_n$, the free group with n generators, the choice that one makes to form a word of length $N + 1$ from a word of length N is independent of either the word or N. Accordingly, the graph of $\mathbf{F}_n$ is a homogeneous tree (of degree $2n$).

Cantor sets of continued fractions

Cantor point sets play an important role in measure theory and in the theory of continued fractions. The Cantor ternary set C is a basic example of an uncountable Borel-measurable set whose measure is zero (see, for example, [5], p. 44 and 63). An important object in the theory of continued fractions is the set $F(n) = \{x \in [0, 1] : x = [0; a_1, a_2, a_3, ...]$ and $a_i \leqslant n$ for all $i\}$, that is, the set of continued fractions of bound n (n being any positive integer). The fact that $F(n)$ is a Cantor point set depends on the property that if

$$x = [0; a_1, ..., a_m, b_{m+1}, b_{m+2}, ...] \quad \text{and} \quad y = [0; a_1, ..., a_m, c_{m+1}, c_{m+2}, ...]$$

are in $F(n)$, then $x < y$ $(x > y)$ if $b_{m+1} < c_{m+1}$ and m is odd (m is even). In particular,

$$\min F(n) = [0; n, 1, n, 1, ...], \max F(n) = [0; 1, n, 1, n, ...]$$

and $F(n)$ can be obtained by first removing from $(0, 1)$ the open intervals

$$(0, [0; n, 1, n, 1, ...]) \quad \text{and} \quad ([0; 1, n, 1, n, ...], 1)\,,$$

then removing the intervals

$$([0; n, n, 1, n, 1, ...]\,,\ [0; n-1, 1, n, 1, n, ...])\,,$$
$$([0; n-1, n, 1, n, 1, ...]\,,\ [0; n-2, 1, n, 1, n, ...])\,,$$
$$..., ([0; 2, n, 1, n, 1, ...]\,,\ [0; 1, 1, n, 1, n, ...])\,,$$

and so on (see [3], p. 971).

A theorem of M. Hall Jr. says that $F(4) + F(4) + \mathbf{Z} = \mathbf{R}$ ([3], theorem 3.1), which is the analogue of $C + C = [0, 2]$. Hall actually proves more general theorems on the nature of $L(A) + L(B)$ for arbitrary Cantor point sets $L(A)$ and $L(B)$. One of the main applications of Hall's theorem is the result that the Markoff spectrum contains every real number greater than 6 (cfr. [1], p. 454). The number 6 has successively been replaced by a best possible value, called Hall's ray (≈ 4.5), by employing a refinement of Hall's original theorem (see [2]).

The set $F(2) + F(2)$ has been used in [4] to prove the existence of certain gaps in the lower Markoff spectrum. It is the proof contained there that originally inspired our geometric construction.

REFERENCES

[1] Cusick, T. W. The largest gaps in the lower Markoff spectrum. *Duke Math. J. 41* (1974), 453-463.

[2] Freiman, G. A. Diophantine approximation and the geometry of numbers (Markov's problem), Kalin. Gosud. Univ., Kalinin, 1975.

[3] Hall, M. Jr. On the sum and product of continued fractions. *Annals of Mathematics, vol. 48, No. 4* (1947), 966-993.

[4] Kinney, J. R. and T. S. Pitcher. On the lower range of Perron's modular function. *Canad. J. Math. 21* (1969), 808-816.

[5] Royden, H. L. *Real Analysis.* 2nd edition, 1968, Collier Macmillan Publishers.

(Reçu le 28 mars 1988)

Marco Pavone

Department of Mathematics
University of California
Berkeley, CA 94720 (USA)

vide-leer-empty