From Venn Diagrams to Peano Curves

LUCIENNE FÉLIX
Paris

Reprinted from "Mathematics Teaching," the Bulletin of the Association of Teachers of Mathematics No. 50, Spring, 1970.

From Venn Diagrams to Peano Curves

LUCIENNE FÉLIX
Paris

We start by translating the structure of a two-valued logic into a Venn diagram, which sets us certain drawing tasks. Gradually losing our original aim, we let these develop and see what happens.

1.

A. In a two-valued logic, the consideration of several independent attributes $\mathscr{A}_1$, $\mathscr{A}_2$, $\mathscr{A}_3$, . . . which are possessed or not possessed by all the elements of a reference set, gives rise to a truth table which can be conveniently laid out in the form of a branching tree, where 'T' stands for 'true' and 'F' for 'false'.

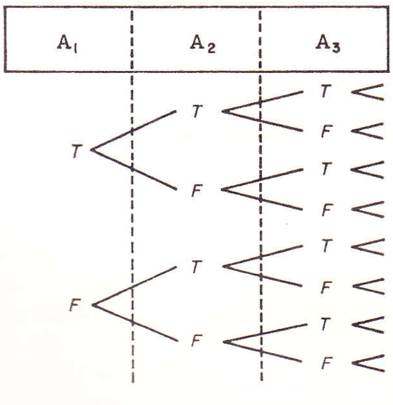


Fig. 1

For two attributes this drawing of a partial order gives four chains forming the sequence TT, TF, FT, FF; for three attributes a sequence of eight chains: TTT, TTF, TFT, TFF, FTT, FTF, FFT, FFF. With n attributes there is a sequence of 2^n chains which exhaust all the possibilities of attributing 'true' or 'false' to the elements of the reference set. So each chain corresponds to a logically defined subset which we will call an *elementary set of order n*. *Fundamental subsets* are those for which the elements give the same response when tested on a single attribute; for example, A_1 is the subset of elements for which $\mathscr{A}_1$ is true; A_1' the subset of elements for which $\mathscr{A}_1$ is false (and so the complement of A_1). Similarly A_2 is the

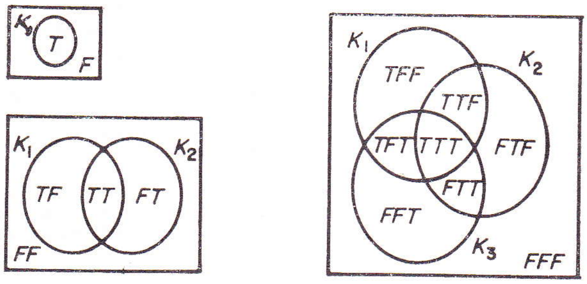


Fig. 2

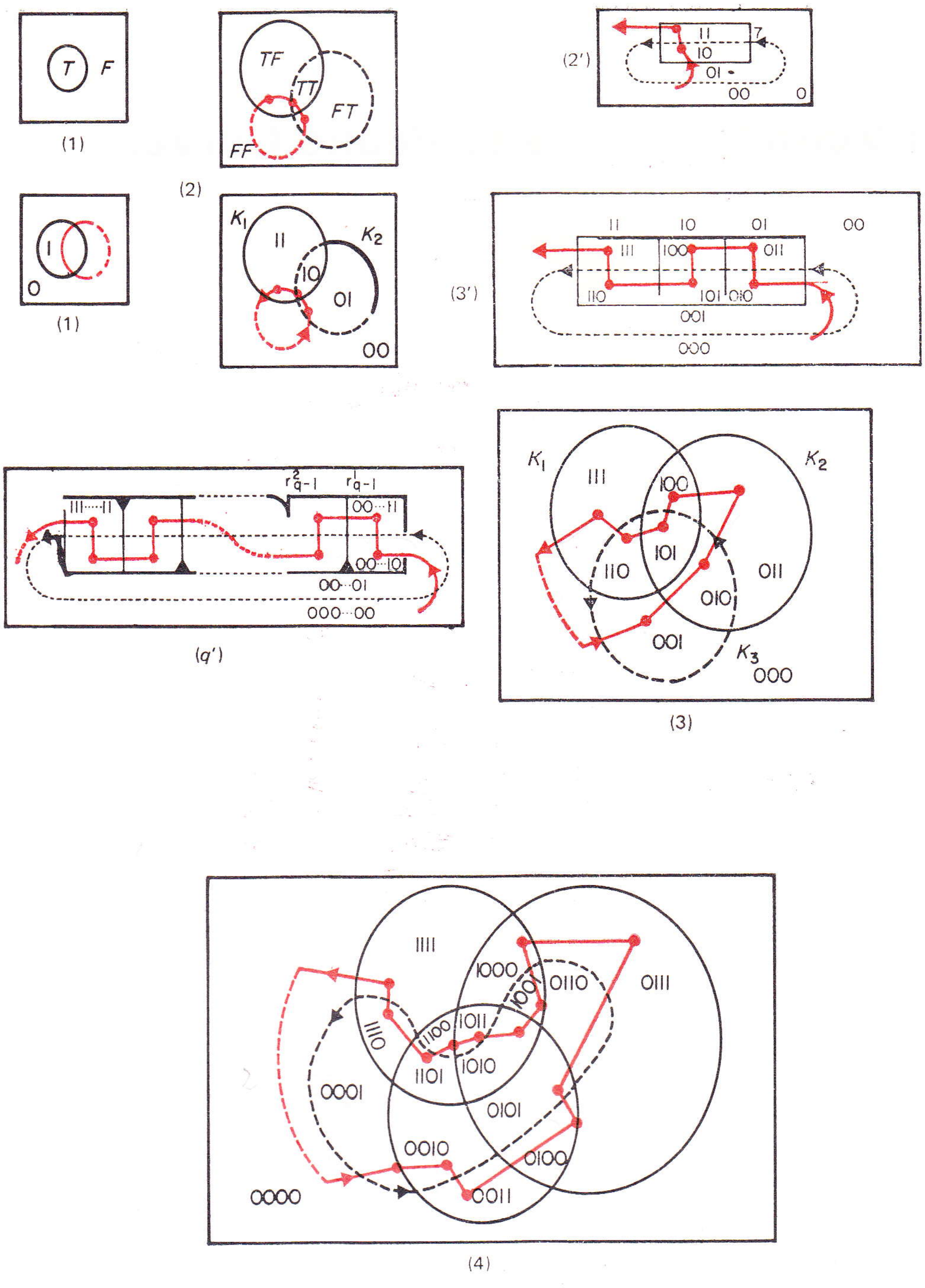

T
F
(1)
TF
TT
FT
FF
(2)
(2')
11
10
01
00
0
1
0
(1)
K1
K2
11
10
01
00
(3')
11
10
01
00
111
100
011
110
101
010
001
000
r_{q-1}^2
r_{q-1}^1
111⋯11
00⋯11
00⋯10
00⋯01
000⋯00
(q')
K1
K2
K3
111
100
101
110
011
010
001
000
(3)
1111
1000
1001
0110
0111
1110
1100
1011
1101
1010
0001
0101
0010
0100
0011
0000
(4)

Fig. 3

subset for whose elements $\mathscr{A}_2$ is true, and so on. Thus A_1 corresponds to the set of chains which have a T in the first place; A_2 the set of chains which have T in the second place, and so on.

B. When there are only a few attributes, the situation described by the tree is represented by a Venn diagram: each fundamental set $A_1, A_2, \ldots$ is represented by a connected region (a single piece) bounded by a simple closed curve (i.e. one which does not cross itself). The set of these curves forms an overlapping pattern in which the parts represent the elementary subsets. *Making the constraints more precise: two boundaries can cut each other, but only in distinct points, and without a common arc between them.* This set of curves is completed by an outer boundary – a rectangle, for example – which is covered by the set of regions at any stage, one of them being a ring.

It is easy to make a drawing for 1, 2 or 3 attributes: we write the members of each chain in each region of the pattern.

But the problem is to determine a strategy in order to carry on the process to 4, 5, . . . attributes. Can we keep to the constraints that we have imposed on ourselves?

To find an answer let us look closely at how we begin.

At the first stage, *Fig.* 3 consists of a loop with its interior marked T and the annular region marked F. At stage (2) we draw the boundary K_2 with a dotted black line to show that it is the last one to be drawn, and write the names of the four chains of the tree. To get to stage (3), we mark a (red) point in each of the elementary regions of stage (2) and join them up, step by step, with red arcs so that each arc crosses the common boundary of the two regions. This arranges the regions in order, and the three non-annular regions form a strip. The linked regions are FT, TT and TF, which is not the natural order of the chains in a tree, where FT and TF would be consecutive.

On the diagram of the strip (2′) (*Fig.* 3), as on the Venn diagram (2), the new order is shown with a directed red arc. By closing this arc in the annular region, we get a new curve K_3 which, drawn as a black dotted line, will serve for the diagram of stage (3). (We must be quite sure that a new boundary does not pass through any point common to the existing curves.)

This is the procedure we follow. To define the procedure more clearly, let us show how to allocate numerals to the $s_{q-1}=2^{q-1}-1$ regions which make up the strip at stage $(q-1)$. As each line K_1 divides each region from the previous stage into two parts, it is natural to use base two notation. In (1) we put 0 for the annular region and 1 for the inside of the loop. In (2) we use the four numerals consisting of two digits. We will agree to mark the annular region 00 and follow the (arbitrary) sense of the red line. At stage (q) we will need 2^q numerals each consisting of q digits.

Suppose the drawing has been completed up to stage $(q-1)$. The diagram (q') of the strip shows us how to pass to stage (q). The strip comprises the successive regions $r^1_{q-1}, r^2_{q-1}, \ldots$ whose numerals, each of $(q-1)$ digits, are

000 . . . 1, 000 . . . 10, 111 . . . 1.

We draw a dotted black line down the middle of the strip; it forms 2^q-2 regions. In order to define the new strip, we give the regions q-digit numerals following the red arc in the order already used. Each elementary region of order $(q-1)$ is divided into two parts which we indicate by writing first 0 and then 1 at the right of the numeral of the region. Therefore the red line alternately crosses a dotted and a full line. In two of the figures we see that an undotted arc is not crossed within the strip, and will not be crossed in what follows: we mark these in thick black. To get a concrete realisation of the strip we must cut the paper along the arcs which form its two edges.

As the q-digit numeral 000 . . . 0 is allocated to the annular region, the numeral 000 . . . 1 is reserved for the region obtained by closing the dotted line; it becomes the first region of the strip at stage (q). The second region of this strip, written 000 . . . 10, is one of the two regions formed in r^1_{q-1} which is linked with the first region.

Lastly, since $s_q=2^q-1$ is clearly odd, the red arc definitely ends in the annular region, and everything is ready for the next stage: the red arc is closed, replaced by a dotted black line, and the programme is re-run.

It is important to stress that each red line, having served to define the strip and its numbering, becomes the boundary at the next stage. The thick black lines along which we can cut become longer and fork in the Venn diagram. In moving from one stage to the next, the digits of each numeral are retained from the left. Consequently the notation will tell us if one region is included in another. The included region has more digits, and so represents a larger number in base two.

We have already pointed out that our numbering does not correspond to the order of the chains in the logical tree. So how can we recognise the elementary region which corresponds to each fundamental set $A_1, A_2, \ldots$? These sets are determined by a certain number of digits from the left. So if q is the number of attributes, A_1 is the set of elementary regions having numerals.

{1}	for $q=1$
{10, 11}	for $q=2$
{100, 101, 110, 111}	for $q=3$

In the same way, 01 . . . represents the set of numerals with left hand digits 01, etc. We can now draw up a useful table which shows the composition of the fundamental regions for each value of q. (We show the natural order of the numerals by an arrow.)

A_1	$\{1 \ldots\}$
	$\uparrow$
A'_1	$\{0 \ldots\}$
A_2	$\{01 \ldots, 10 \ldots\}$
	$\uparrow \longrightarrow \downarrow$
A'_2	$\{00 \ldots, 11 \ldots\}$
A_3	$\{001 \ldots, 010 \ldots, 101 \ldots, 110 \ldots\}$
	$\uparrow \longrightarrow \downarrow \longrightarrow \uparrow \longrightarrow \downarrow$
A'_3	$\{000 \ldots, 011 \ldots, 100 \ldots, 111 \ldots\}$
A_4	$\{0001 \ldots, 0010 \ldots, 0101 \ldots, \ldots, 1101 \ldots, 1110 \ldots\}$
	$\uparrow \longrightarrow \downarrow \longrightarrow \uparrow \qquad \longrightarrow \uparrow \longrightarrow \downarrow$
A'_4	$\{0000 \ldots, 0011 \ldots, 0100 \ldots, \ldots, 1100 \ldots, 1111 \ldots\}$

The law is now obvious. To every Boolean function formed by using complementation, intersection and union, corresponds a set of numerals which can easily be written down from our knowledge of how to represent inclusion.

C. Rectilinear diagrams

Suppose we regard the numerals we have used as sets of digits following a point: we have numbers written in base two with the integral part zero. All of these n-digit numbers, for all values of n, belong to the interval $I=[0, 1]$. Knowing the first digits after the point allows us to associate with each elementary region an interval which is closed on the left and open on the right. The union of these intervals is $[0, 1]$. Sets of these intervals are associated with the fundamental subsets. Let us mark the images of $A_1, A_2, \ldots$ in red, and the images of their complements in black. Then a sequence of elementary regions from successive stages, associated with the numbers $0{\cdot}a_1$, $0{\cdot}a_1a_2$, $0{\cdot}a_1\,a_2\,a_3, \ldots$ $(a_i \in \{0, 1\})$, corresponds to a sequence of nested intervals.

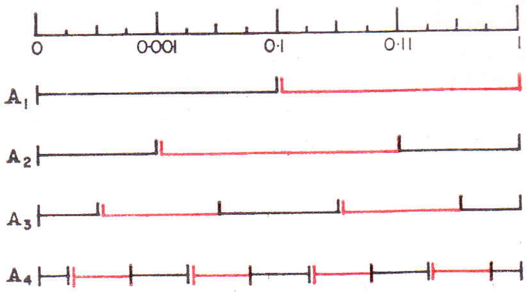


Fig. 4

If we imagine an infinite sequence of attributes, an elementary region becomes associated with a real number, and so with a point of [0, 1]. Conversely, each point of the interval can be associated with at least one infinite sequence of elementary regions, each one included in the next. It is necessary to distinguish between, say, 0·110000 . . . , and 0·101111 . . . , which are different forms of the same number 0·11, because they do not represent the same region. In the strip, however, at each stage, the corresponding regions are consecutive, for example, in (2) we see 11 and 10; in (3) we see 110 and 101; and in (4) we see 1100 and 1011, etc.

The question which now emerges is whether each sequence of regions, each included in the next, can be considered as defining a limiting region. We will see how these conditious are met in the next section.

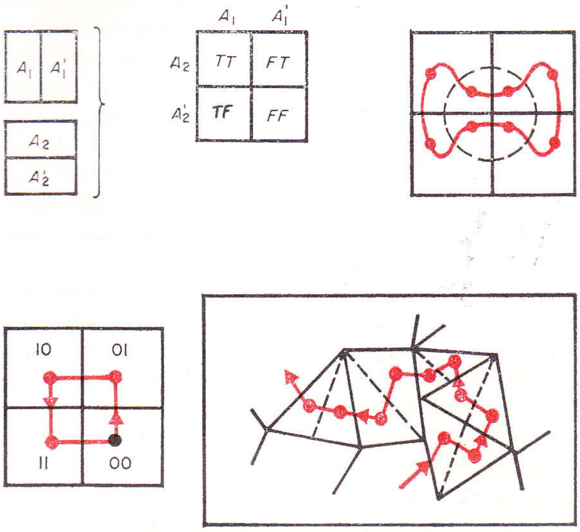


Fig. 5

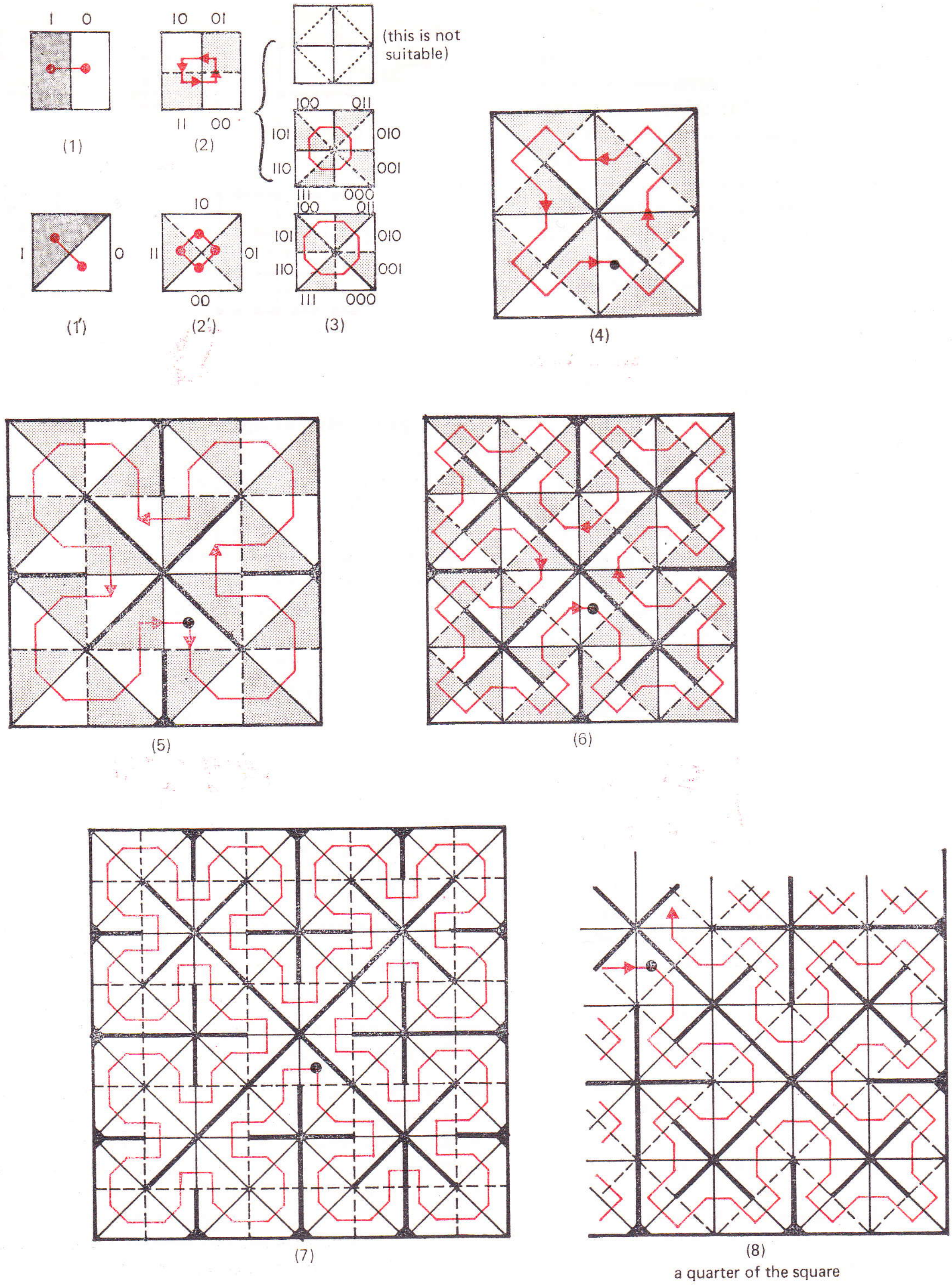


Fig. 6

2. Dissection diagrams

In honour of Lewis Carroll we often draw 'Carroll diagrams' which, when there are two attributes, represent each of the fundamental subsets A_1 and A_2 by a half of a square formed by bisecting opposite sides. By putting one on top of the other, we get a perfect representation of the tree of order two. Unlike the Venn diagram, the Carroll diagram allows each fundamental region to play the same role as its complement. There is no precedence accorded to the values 'true' and 'false'.

How can we develop the diagram, though, to deal with more than two attributes? It is tempting to represent A_3 by the interior of a circle whose centre is at the centre of the square, and continue with simple closed curves. But this is to recreate the Venn diagram from an unsuitable beginning. We have to try another direction. It is essential always to divide each of the regions at stage $(q-1)$ into two since we are dealing with a two-valued logic. To help separate the regions we will still preserve a cyllic order and, naturally, keep to a base two notation. But the big difference is that the red line (L_1), which shows the order of the regions and which therefore defines a strip, will no longer act as a boundary. The boundaries will be found by dissecting the square into elementary rectangles as we please.

In the diagrams, if the dissection corresponding to the order $(q—1)$ is marked with black lines and the edge of the strip with thick black lines, the next order is obtained by tracing with a black dotted line a set of boundary arcs that have been attached to preceding arcs. Then in order to change the numeration from order $(q-1)$ to order q, we draw the directed red line which crosses dotted and full arcs alternately but does not cross the cut arcs (thick black). The red line joins up with itself: its first and last points coincide.

As we are using polygonal dissections, we will make the lines L_1 the lines whose vertices will best lead to an elegant design: centres of squares or rectangles, point of intersection of the medians of triangles.

But we still have to choose a dissection which will work. We will not write the well-known numerals this time. The elementary subsets at stage (q) of the strip are naturally numbered successively odd and even. The fundamental subset A_q comprises the regions whose numerals have the same parity: the odd numbers, for example. They are shaded in the first stage in *Fig.* 6.

Dissection into right-angled isosceles triangles

A particularly elegant figure is obtained by dividing the given square by a diagonal, and then following the rule: each triangle in stage $(q-1)$ is divided into two triangles in stage (q) by an altitude. It is also possible to start with the four-square Carroll diagram, which becomes the same as the above at stage (3).

Dissection into rectangles and squares

We obtain squares and rectangles alternately (see *Fig.* 7). We do not use parallel strips since the need to connect them up would necessitate working on the surface of a cylinder! So we alternate the parallels to the two directions of the square.

By drawing figures with our conventions we come up against an impossibility at stage (5). Another false trail is shown in (3′) and (4′).

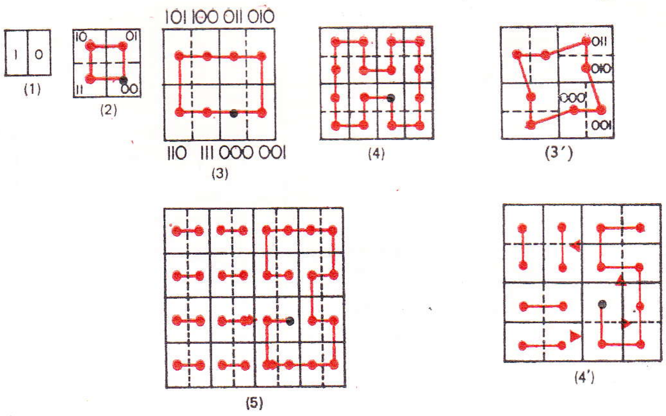


Fig. 7

3. A geometrical approach

A. We abandon the two-valued logic approach in order to study the dissections and the sequences L_1 which will cover the square.

Instead of splitting each square into two and then two again, we will split it into four so that we can choose between the two contours (α) and (α') which are now equally valid.

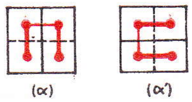
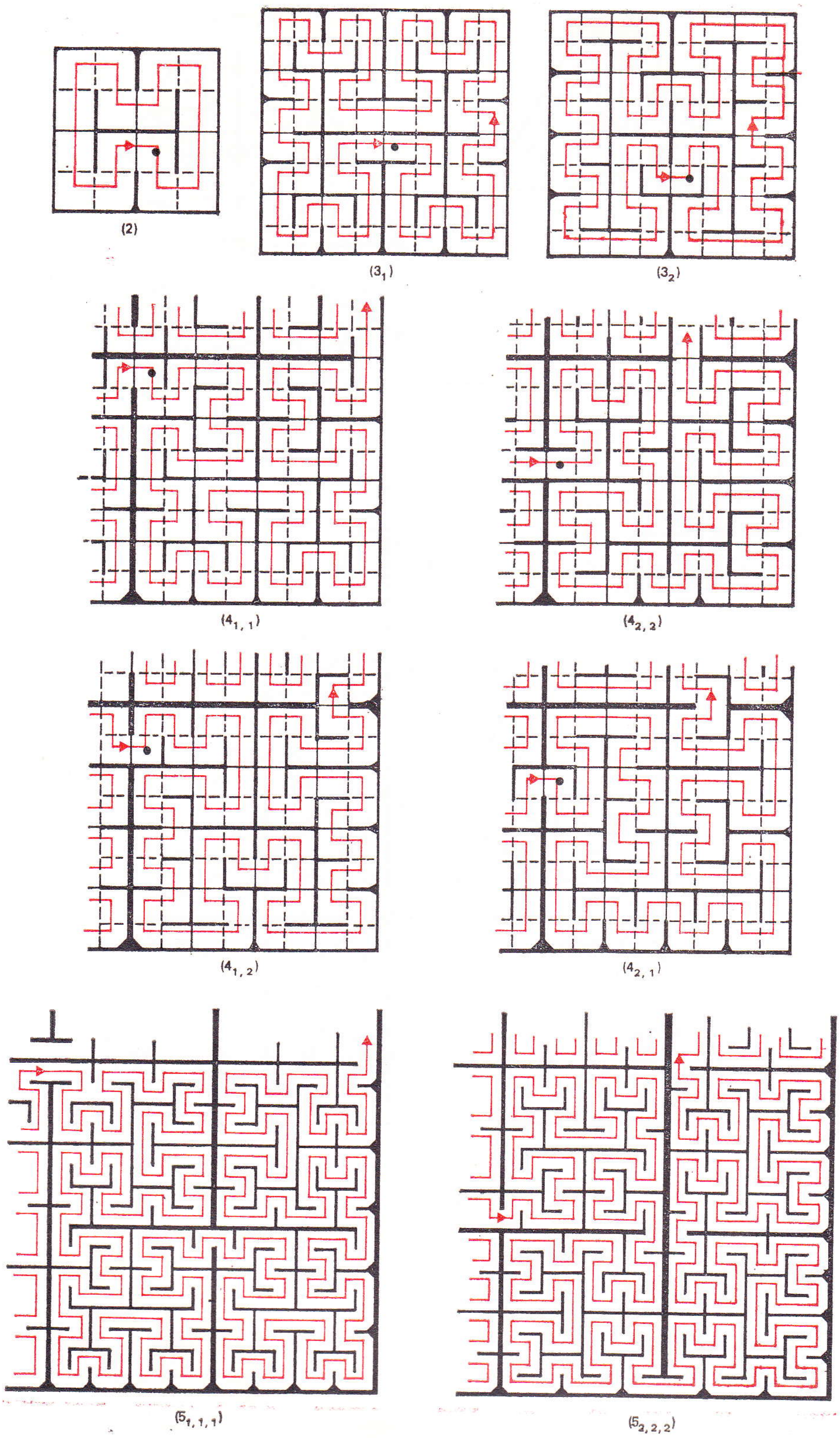


Fig. 8

The best notation, naturally, will be base four. At each stage the diagram is determined by choosing a starting point. From (3), for example, this freedom leads to two different drawings at each step from (q) to $(q+1)$. We can best take account of this by noticing the form of the cut lines (thick black); this is why, in the absence of a theoretical study, we must continue the graphical study a little further. It is only in drawing the figures that it also becomes clear how hypotheses about the nature of the junctions intervene.

B. Since division into two parts no longer holds, we may start to consider base three, which naturally leads to triangular dissections. In order to work in base three, the surface D which we cover with the strips leading to the curves L_1 will not be a square but an equilateral triangle (since we will only consider straightforward symmetries). Each equilateral triangle will be decomposed by radii of the circumscribed circle into three congruent triangles with angles 120°, 30°, 30°, and each of these will be decomposed by the trisectors of the obtuse angle into an equilateral triangle and two triangles similar to the original. This will work since we can satisfy the conditions of linking. The construction is now easy.

The number of regions in the strip at stage (q) is 3^q this time, and this is also the number of vertices on the line L_q.



$(4_{1,1})$, $(4_{2,2})$, $(4_{1,2})$, $(4_{2,1})$, $(5_{1,1,1})$ and $(5_{2,2,2})$ show a quarter of the square

Fig. 9

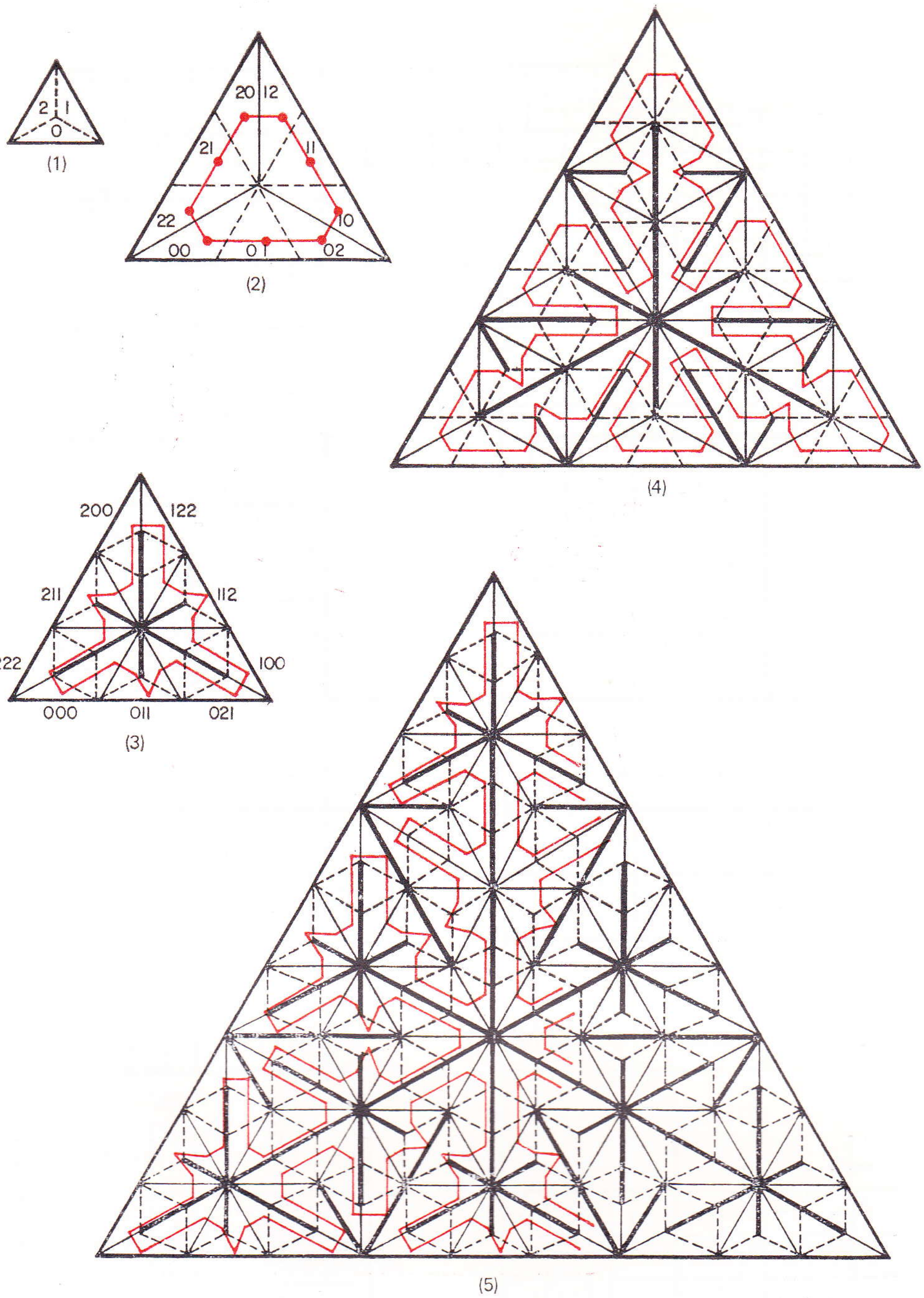


Fig. 10

In each case, whatever base we use, the notation will lead to a rectilinear representation, as we showed in the case of the Venn diagram. The numbers are always understood as a set of digits following the point, the integral part being zero. We are always working therefore in the interval $I=[0, 1]$.

4. Peano curves

The sequences of lines L_q corresponding to our dissection has a limit when q tends to infinity. A limiting curve L of this kind is a Peano curve. We will show that it exists and exhibit some of its properties.

A. Let b be the base of numeration (we have already used 2, 3 and 4). Each strip B_q is formed from b^q elementary regions covering the initial domain D. The number of regions tends to infinity with q. At the same time, the greatest diameter of these regions, say d_q, tends to zero.

(1) Let there be q digits after the point: $t_q=0{\cdot}a_1a_2a_3 \ \ldots$, where t_q is written in the chosen base. The number with q digits is the numeral

attached to one of the elementary regions, and *vice-versa*. Each region can be defined as the intersection of the regions represented by the sequence obtained at previous stages of the numeral: $0{\cdot}a_1$, $0{\cdot}a_1a_2$, $0{\cdot}a_1a_2a_3$, . . . This is a nested sequence of regions.

(2) Now consider a real number belonging to the interval I; i.e. a number defined by an unlimited sequence of digits a_1. The infinite sequence of nested regions whose diameter tends to zero has a limit point m of D. Projecting onto two coordinate axes, this point is given by its coordinates x and y which are the limits of the nested segments obtained by projection of the elementary regions.

(3) Let t be any number of $[0, 1]$. We associate with it a point m of D defined by a function $f{:}I{\to}D$. The fact that the numbers which can be written with a finite number of digits have two infinite forms does not matter since the two versions correspond to neighbouring regions and so lead to the same limit. (For example, notice that on the base two diagram, $0{\cdot}11=0{\cdot}11000\ldots=0{\cdot}10111\ldots$)

But is the converse true, that a point m of D corresponds to a given number t? The function f is obviously surjective since the strips, considered as composed of edges, cover D. But some points m clearly correspond to several numbers because of the slits. A point on a slit is the limit of points which are not adjacent on the strip, and so are not adjacent in the notation. If we look at the growth of the slits in our drawings, which we can do because we have looked at several stages, we see that a part can correspond, depending on the dissection, to 1 or 2 or 3 or 4 numbers, or even 6 in the case of *Fig.* 10.

Therefore the function is not bijective.

But *the function is continuous*, for by the construction, every two neighbouring numbers t have neighbouring images m. Formalising this: let m_0 be the image of t_0. To show that dist $(mm_0) < d$ it is sufficient, provided q is chosen so that $d_q < d$, to note that t has more than q digits in common with t_0.

B. The limit of the sequence L_i

We choose two axes and use cartesian coordinates (x, y) which define the point m. Each curve L_q is the set of points obtained from t by two functions

$$g_q: t \to x \qquad h_q: t \to y$$

These functions are continuous since the line L_q is. They are defined on I. We note that in order to obtain the best expressions for these functions we do not choose the vertices of L_q which we used to get the best diagrams. But since the curve L_q is defined by arcs determined by a particular starting point, it is obvious that it will not be easy to find expressions for the functions.

We consider the continuous functions g_q defined on I. They form a sequence. We show that when q tends to infinity this sequence yields a limit function g.

For any $d > 0$ we can choose q large enough to make $d_q < d$. Then for any $t \in I$, and any q_1, q_2 greater than q,

$$| g_{q_1}(\in) - g_{q_2}(t) | < d.$$

This shows that the sequence of functions tends uniformly to a limit g defined and continuous on I.

In the same way, the functions h_q have a limit h defined and continuous on I. Consequently the sequence L_i has a limit, the curve L, which is the set of points satisfying $x=g(t)$, $y=h(t)$. (The curve becomes a trajectory if t is taken to be time.)

So we obtain the Peano curve corresponding to each of our dissections. This curve, passing through all points of the domain D, is the image of I, and hence of a line segment, by the function f; it is defined on I, surjective and continuous, but not bijective. Since the curve obviously has no tangents, the functions g and h are continuous but not differentiable.

5. A historical note

The definitions of the functions g_q and h_q, and then of g and h, can only be expressed algebraically by starting from the chosen base of numeration. Peano made this clear in a short note published in 1890 (*Math. Ann.*, Vol. 46). The curve he defined, without using any geometry or calculations, is not one which can be derived from taking to the limit any of the sequences we have used. In effect he used base three and filled a square with a curve which is not closed. (This latter point is not significant since we can obtain a closed curve by applying symmetry operations.)

The formulae are extraordinarily simple and can be written in a few lines with modern notation.

We write $t = 0{\cdot}a_1a_2 \ldots a_n \ldots$ (base 3)

$x = 0{\cdot}b_1b_2 \ldots b_n \ldots$

$y = 0{\cdot}c_1c_2 \ldots c_n \ldots$

and $a_2+a_4+\ldots+a_{2n}\equiv\alpha_n \pmod 2$

$a_1+a_3+\ldots+a_{2n-1}\equiv\alpha_n' \pmod 2$.

A permutation p of $\{0, 1, 2\}$ is defined by $p(0)=2$, $p(1)=1$, $p(2)=0$.

The formulae are then

$$b_1=a_1,\ b_n=p^{\alpha_{n-2}}(a_{2n-2}),\ c_n=p^{\alpha^1_n}(a_{2n}).$$

But Peano said nothing which leads to these laws. The following year, Hilbert, in the same periodical, showed a construction very similar to our *Fig.* 9 (the curve not being closed). He used base six notation by taking only successive numerals independently of the structure of the diagram, and by introducing the projections x and y. Although giving a more intuitive geometrical example than the arithmetical example used by Peano, he failed to show the processes which led up to the example.

Cantor had already given examples of a bijective correspondence between points of a line segment and a region, but his functions were discontinuous. The object of Peano's note was to obtain continuity, but he achieved it at the expense of bijectivity.